Oregon State University

COLLEGE OF ENGINEERING

On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data

Mark Roth

Oregon State University

COLLEGE OF ENGINEERING

On the Role of Spatial Clustering Algorithms in Building Species Distribution Models from Community Science Data

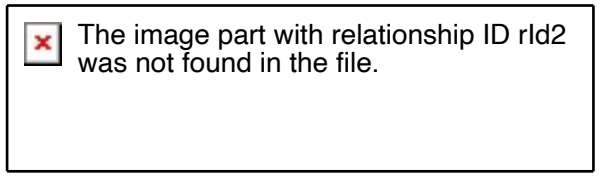
Mark Roth[1], Dr. Tyler Hallman[2],

Dr. W. Douglas Robinson[3], Dr. Rebecca Hutchinson[1,3]

1: Department of Electrical Engineering & Computer Science, Oregon State University

2: Swiss Ornithological Institute, Sempach, Switzerland

3: Department of Fisheries, Wildlife, & Conservation Sciences, Oregon State University



Overview

- Motivation & Background
  - Biodiversity Loss, Species Distribution Models, Community Science, Occupancy Models
- Research area
  - **Site Clustering Problem**
- Experiments, Results, Contributions
- Future Work

Oregon State University
College of Engineering

Biodiversity

"Biodiversity is the foundation of [our] social and economic systems, yet we have not managed to solve the extinction crisis"

– Leah Gerber, Director of the Center for Biodiversity Outcomes at Arizona State University

https://www.nytimes.com/2021/09/30/opinion/animal-extinction.html

Oregon State University
College of Engineering

Biodiversity Loss

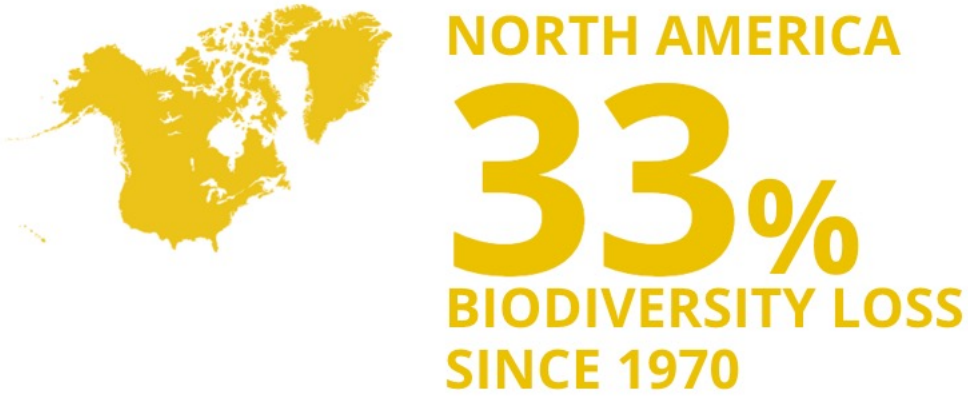
- The world has seen a 68% decline in animal populations since 1970

https://www.worldwildlife.org/magazine/issues/summer-2021/articles/a-warning-sign-where-biodiversity-loss-is-happening-around-the-world

Oregon State University
College of Engineering

Biodiversity Loss

- The world has seen a 68% decline in animal populations since 1970



https://www.worldwildlife.org/magazine/issues/summer-2021/articles/a-warning-sign-where-biodiversity-loss-is-happening-around-the-world

Oregon State University
College of Engineering

Biodiversity Loss

- The world has seen a 68% decline in animal populations since 1970

NORTH AMERICA

33%

BIODIVERSITY LOSS
SINCE 1970

LATIN AMERICA AND
THE CARIBBEAN

94%

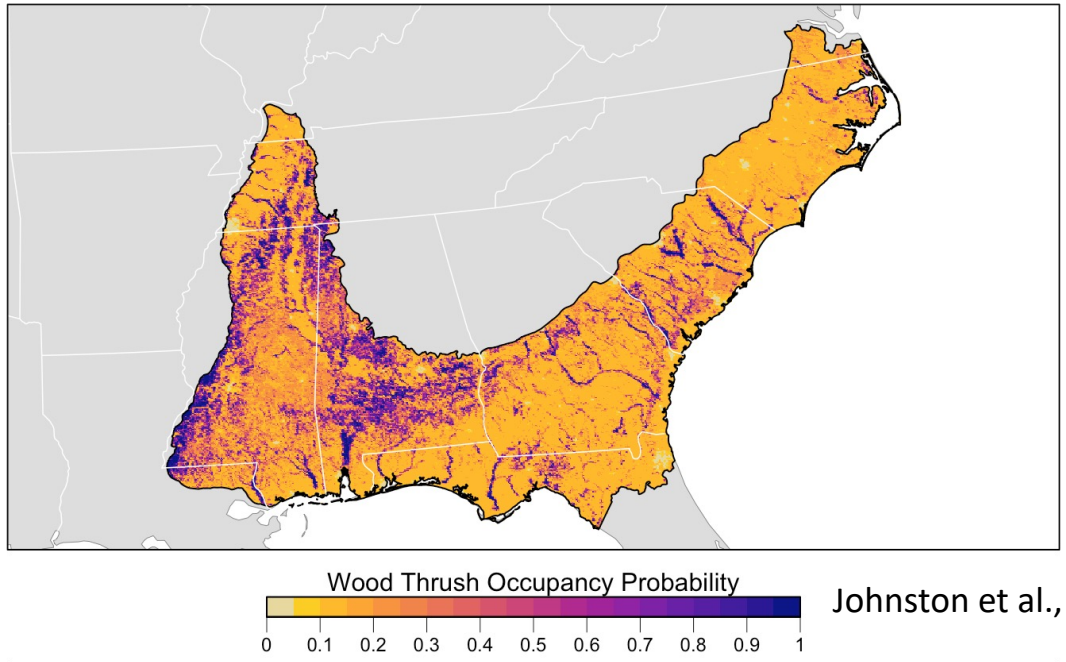
BIODIVERSITY LOSS
SINCE 1970

https://www.worldwildlife.org/magazine/issues/summer-2021/articles/a-warning-sign-where-biodiversity-loss-is-happening-around-the-world

Oregon State University
College of Engineering

Species Distribution Models (SDMs)

- Tools that predict the pattern of species activity
  - Integral in designing solutions to support threatened species



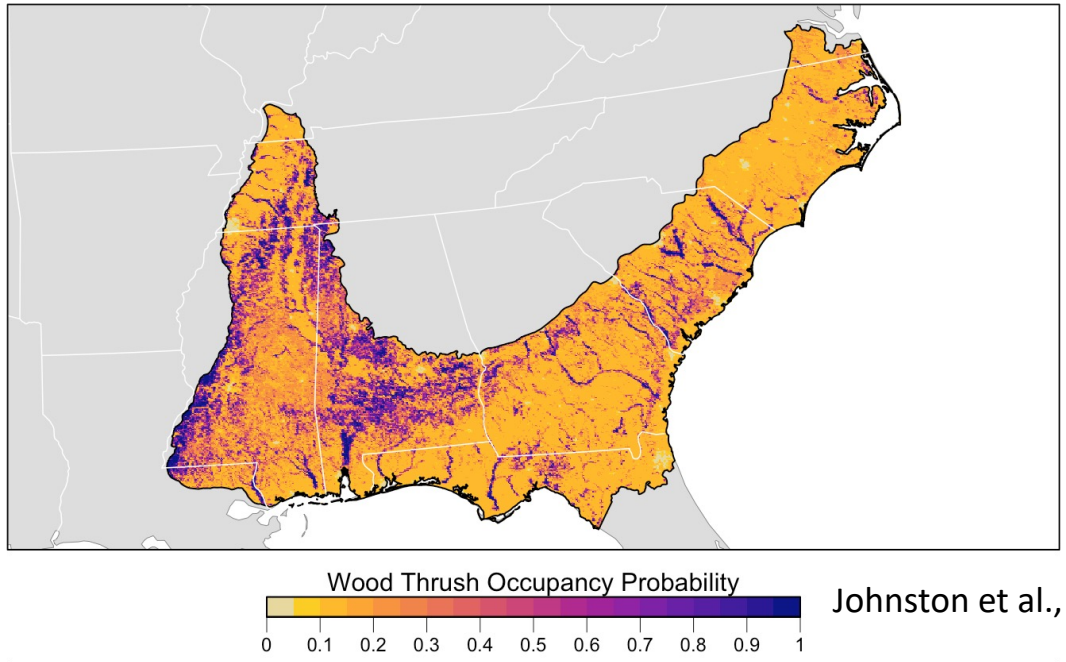
Johnston et al., 2021

Oregon State University
College of Engineering

Species Distribution Models (SDMs)

- Tools that predict the pattern of species activity
  - Integral in designing solutions to support threatened species

- Extent and accuracy of SDMs depend on the range and quality of the biodiversity dataset



Johnston et al., 2021

Oregon State University
College of Engineering

Community Science (also known as citizen science)

- Voluntary crowdsourced data collection

Oregon State University
College of Engineering

Community Science (also known as citizen science)

- Voluntary crowdsourced data collection
- Low barriers to contribute

Oregon State University
College of Engineering

Community Science (also known as citizen science)

- Voluntary crowdsourced data collection
- Low barriers to contribute
- Growing in volume and quality

Oregon State University
College of Engineering

Community Science (also known as citizen science)

- Voluntary crowdsourced data collection
- Low barriers to contribute
- Growing in volume and quality

- We can construct spatially and temporally comprehensive SDMs !

Oregon State University
College of Engineering

Imperfect Detection

- Probability of detecting a species given that it is present is less than 1

Oregon State University
College of Engineering

Imperfect Detection

- Probability of detecting a species given that it is present is less than 1

- Ignoring imperfect detection can lead to biased estimates of occupancy (Guillera-Arroita et al., 2014)

Oregon State University
College of Engineering

Imperfect Detection

- Probability of detecting a species given that it is present is less than 1
- Ignoring imperfect detection can lead to biased estimates of occupancy (Guillera-Arroita et al., 2014)
- Occupancy Models!

Oregon State University
College of Engineering

Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:

Oregon State University
College of Engineering

Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives

Oregon State University
College of Engineering

Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives
  2. N observations are organized into a set of ≤N *sites*

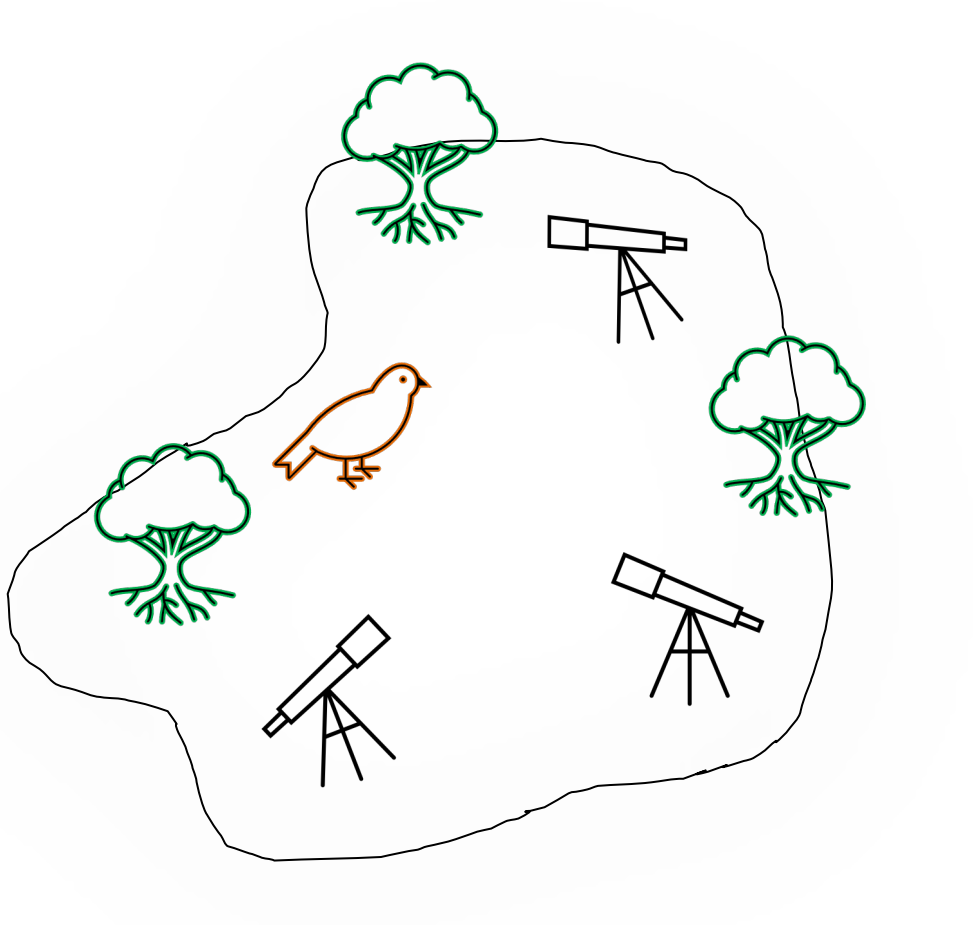
Oregon State University
College of Engineering

Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives
  2. N observations are organized into a set of ≤N *sites*
  3. At each site, we assume *closure*: the occupancy status remains unchanging across all observations

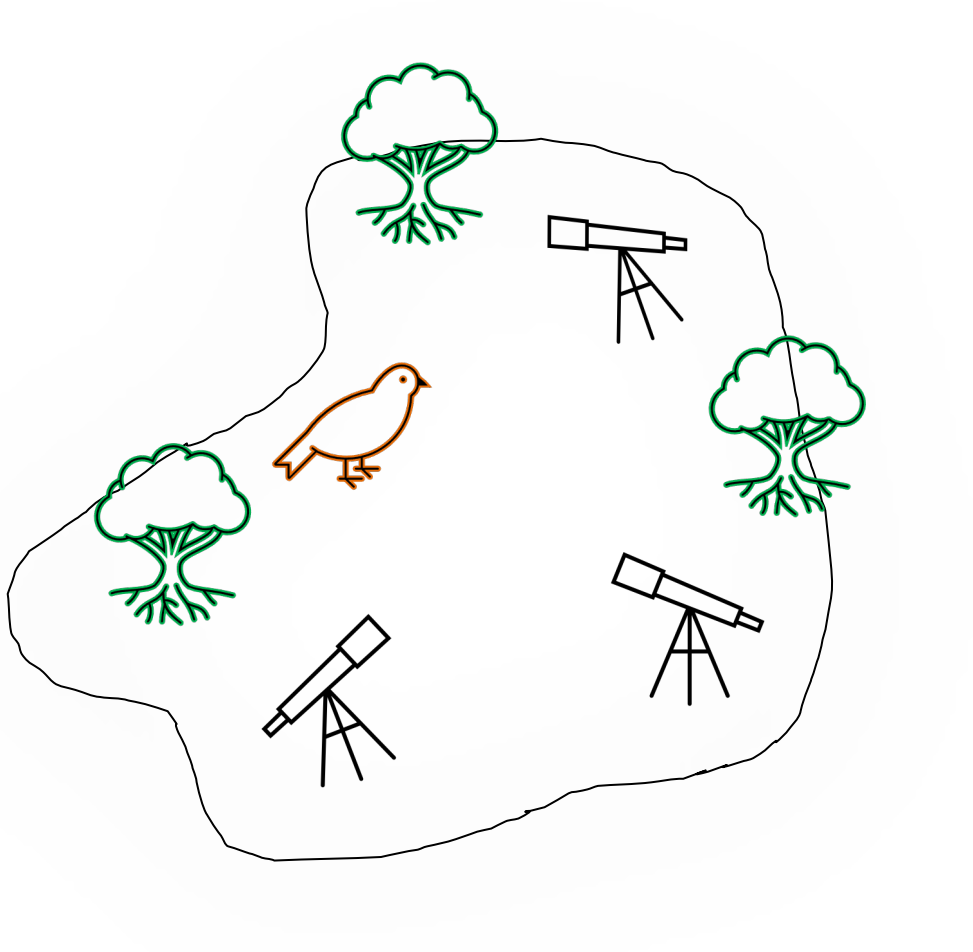
Oregon State University
College of Engineering

Occupancy Model – Intuition



Oregon State University
College of Engineering

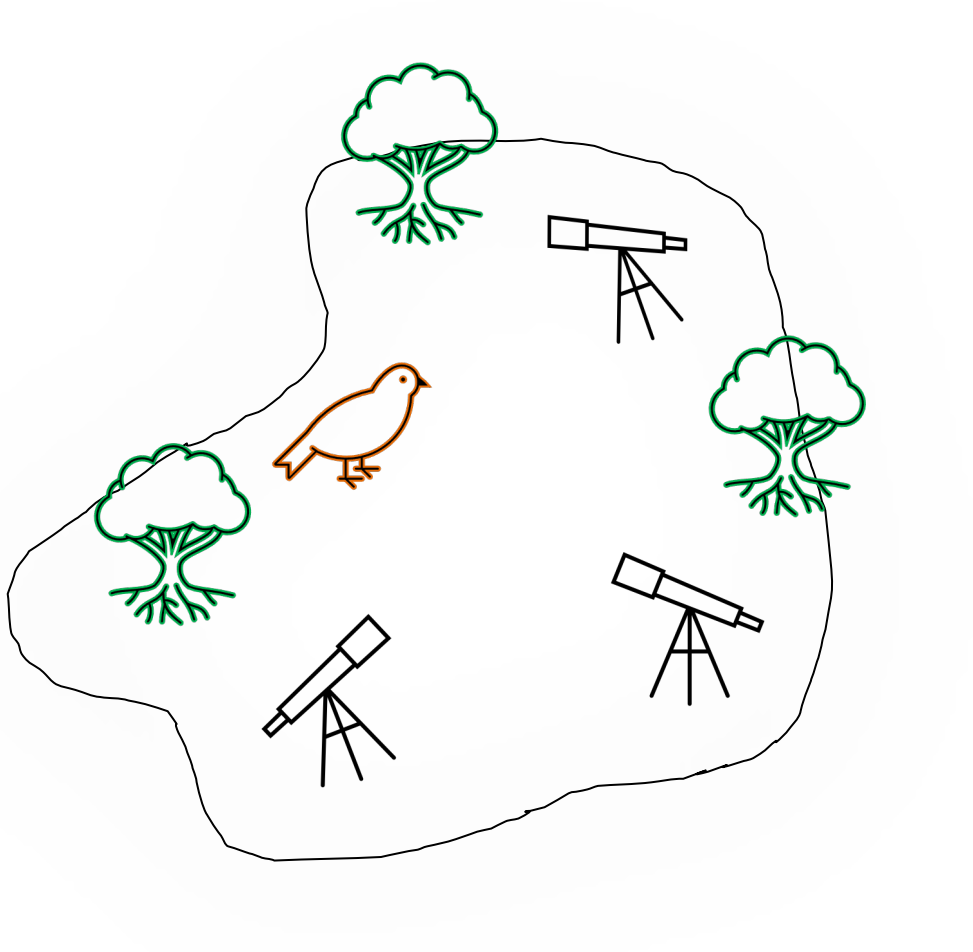
Occupancy Model – Intuition



Observations: [0, 0, 1]

Oregon State University
College of Engineering

Occupancy Model – Intuition

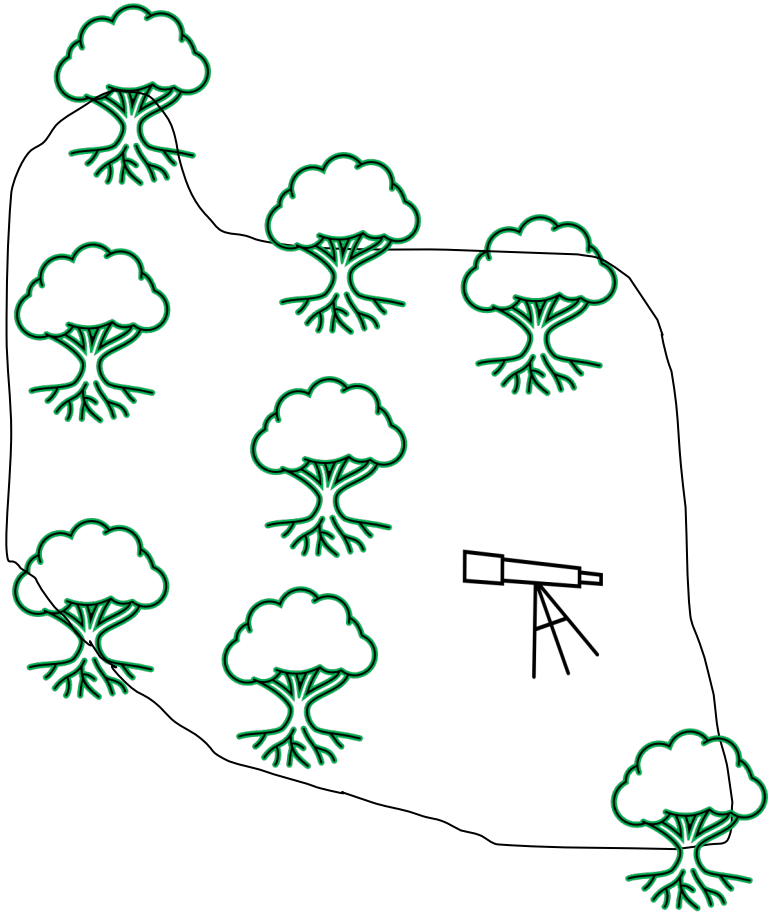


Observations: [0, 0, 1]

Detection probability = 1/3

Oregon State University
College of Engineering

Occupancy Model – Intuition



Oregon State University
College of Engineering

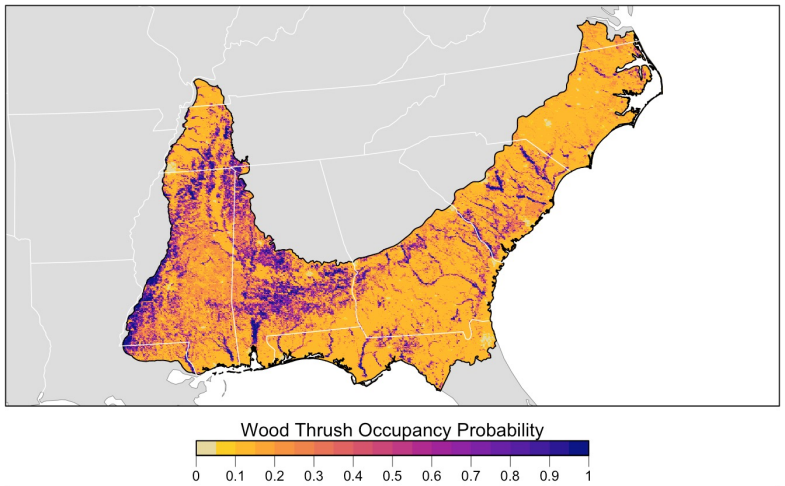
Occupancy Model Output

- Provides equations for occupancy & detection probabilities

Oregon State University
College of Engineering

Occupancy Model Output

- Provides equations for occupancy & detection probabilities
  - occ prob $= o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
  - det prob $= d\_var_0 + .84 * d\_var_1 + .66 * d\_var_2$



Oregon State University
College of Engineering

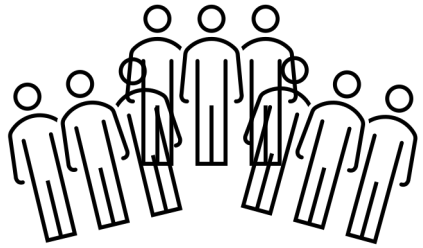
Occupancy Models

- Rely on a few key assumptions to account for imperfect detection:
  1. No false positives
  2. N observations are organized into a set of ≤N *sites*
  3. At each site, we assume *closure*: the occupancy status remains unchanging across all observations

Scientists design sites prior to sampling to ensure closure, but this is not the case with community science!

Pathway to mitigate biodiversity loss

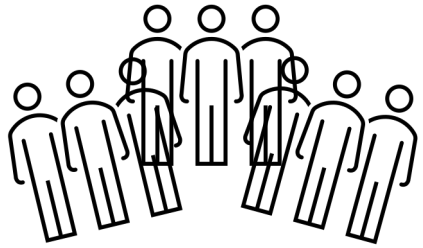
Oregon State University
College of Engineering



Unstructured, crowdsourced biodiversity datasets; imperfect detection

Pathway to mitigate biodiversity loss

Oregon State University
College of Engineering

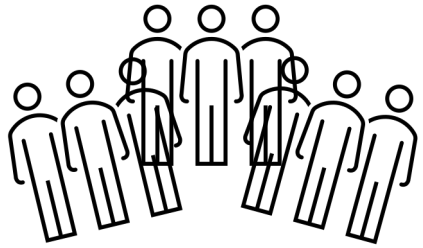


→ OMs

Unstructured, crowdsourced biodiversity datasets; imperfect detection

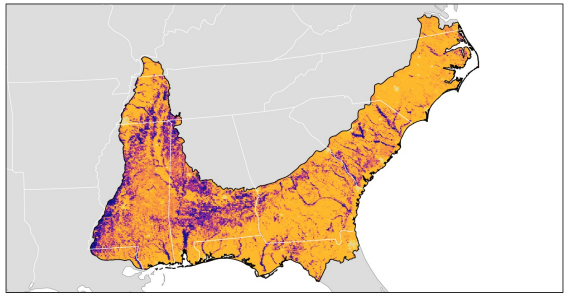
Pathway to mitigate biodiversity loss

Oregon State University
College of Engineering



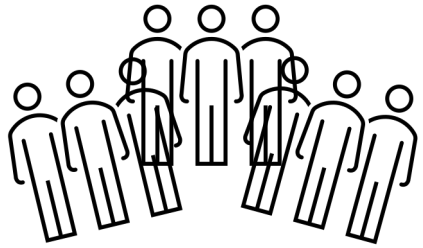
Unstructured, crowdsourced biodiversity datasets; imperfect detection

→ OMs {



Pathway to mitigate biodiversity loss

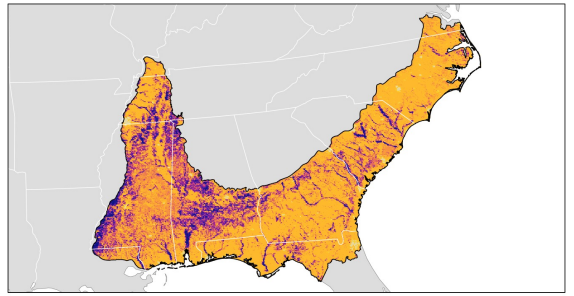
Oregon State University
College of Engineering

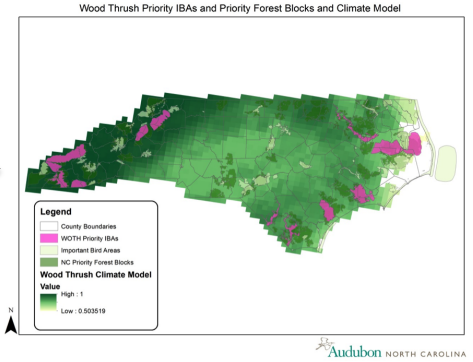


Unstructured, crowdsourced biodiversity datasets; imperfect detection

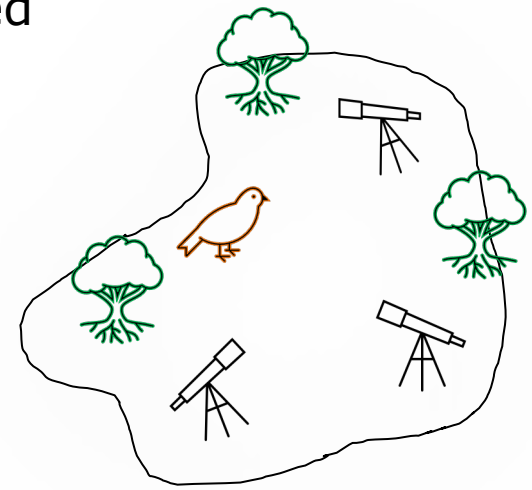
Group observations into sites while maintaining closure

OMs





Natural Resource Management



Group observations into sites while maintaining closure

Oregon State University
College of Engineering

Site Clustering Problem

Group observations into sites while maintaining closure

Oregon State University
College of Engineering

Site Clustering Problem

Oregon State University
College of Engineering

Group observations into sites while maintaining closure

1. Discover the optimal number of sites automatically

Site Clustering Problem

Oregon State University
College of Engineering

Group observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior

Site Clustering Problem

Oregon State University
College of Engineering

Group observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior
3. Consider similarity in geospatial & feature space

Site Clustering Problem

Oregon State University
College of Engineering

Group observations into sites while maintaining closure

1. Discover the optimal number of sites automatically
2. Respect geospatial & temporal constraints imposed by species behavior
3. Consider similarity in geospatial & feature space
4. Run efficiently on large datasets

eBird

Oregon State University
College of Engineering

The Cornell Lab of Ornithology

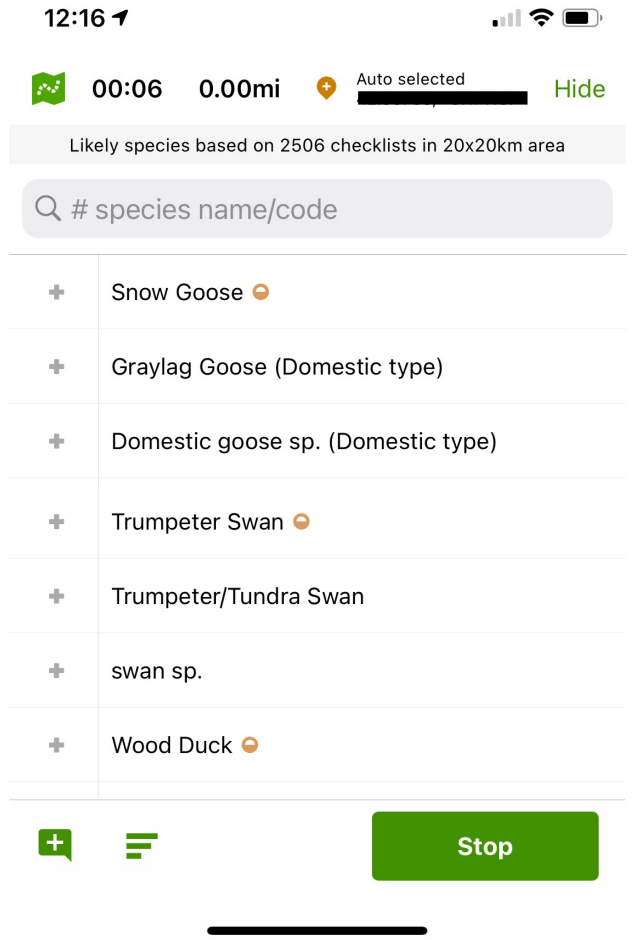
Western Tanager (WETA), eBird.org

Hermit Warbler (HEWA), eBird.org

Oregon State University
College of Engineering

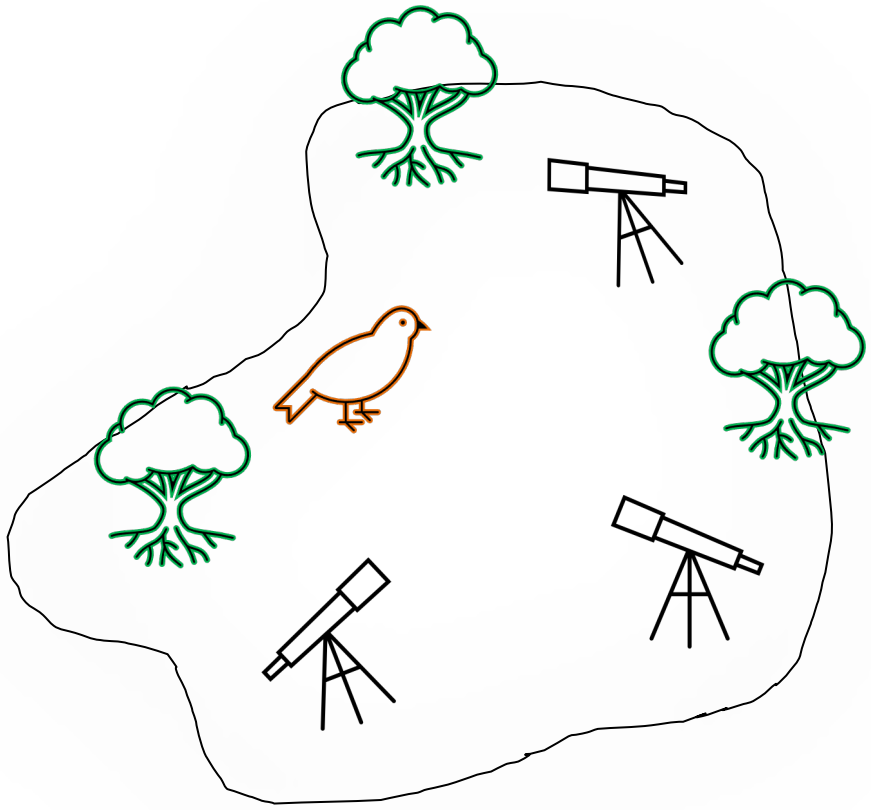
The eBird Community Science Program

- Created in 2002
- Mobile application to record birding observations (checklists)
- >1 billion checklists submitted



Oregon State University
College of Engineering

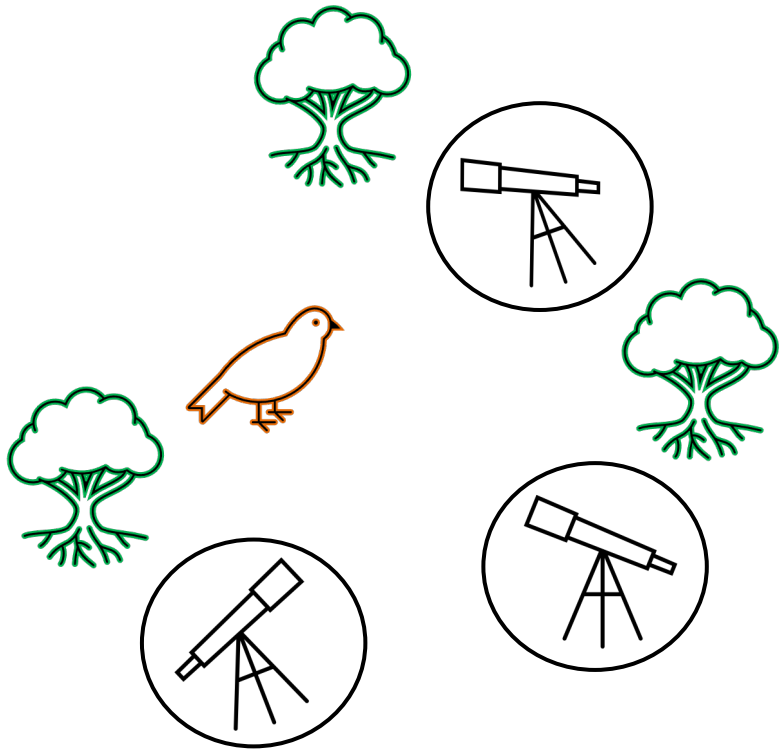
Existing Methods



Oregon State University
College of Engineering

Existing Methods

1. eBird Best Practices
   - Same observer, same latitude-longitude coordinate, > 1 visit and at most 10 visits



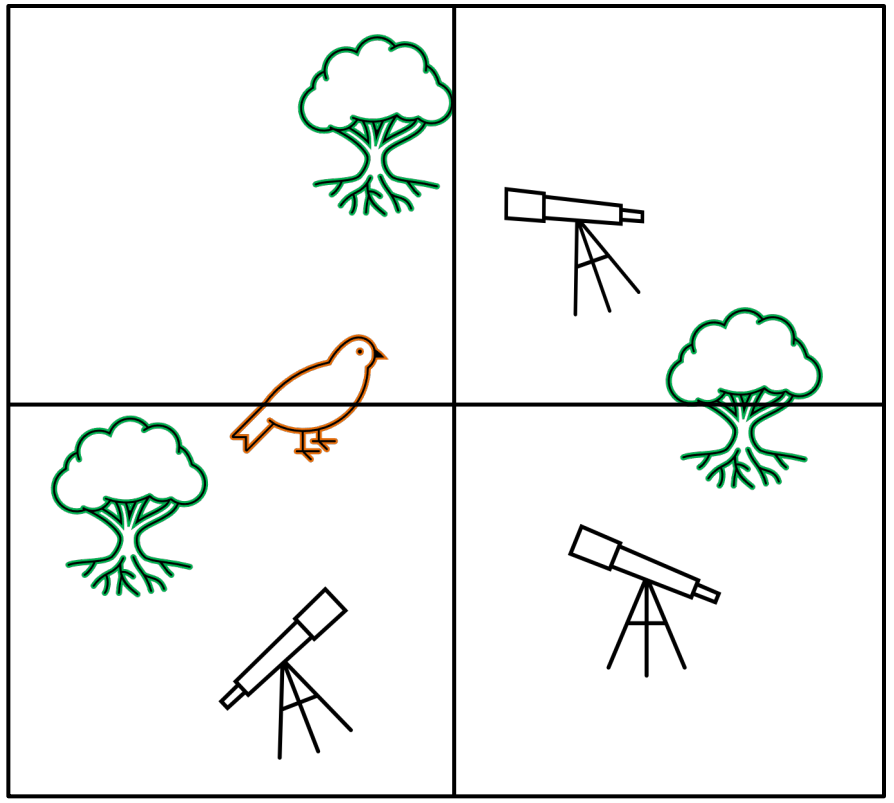
Oregon State University
College of Engineering

Existing Methods

1. eBird Best Practices
   - Same observer, same latitude-longitude coordinate, > 1 visit and at most 10 visits

Retains less than 25% of available data!

2. Grid
   - Most commonly, 1x1km



- Can we improve upon the existing methods by framing the **Site Clustering Problem** as a spatial clustering problem?

Oregon State University
College of Engineering

Existing Spatial Clustering Algorithms

Existing Spatial Clustering Algorithms

- k-means (Llyod, 1982)
- CLARANS (Ng & Han, 2002)

] Partitioning

Oregon State University
College of Engineering

Existing Spatial Clustering Algorithms

- k-means (Llyod, 1982)
- CLARANS (Ng & Han, 2002)

] Partitioning

- DBSCAN (Ester et al., 1996)
- DBRS (Wang & Hamilton, 2003)

] Density Based

- SKATER (Assunção et al., 2006)
- REDCAP (Guo, 2008)

] Regionalization

- For a more complete review see Liu et al. 2012

Oregon State University
College of Engineering

Algorithms in this report

- lat-long
- rounded-4

Oregon State University
College of Engineering

Algorithms in this report

- lat-long
- rounded-4
- Density-based spatially-constrained (DBSC) (Liu et al., 2012)

Oregon State University
College of Engineering

Algorithms in this report

- lat-long
- rounded-4
- Density-based spatially-constrained (DBSC) (Liu et al., 2012)
- clustGeo (Chavent et al., 2018)

Oregon State University
College of Engineering

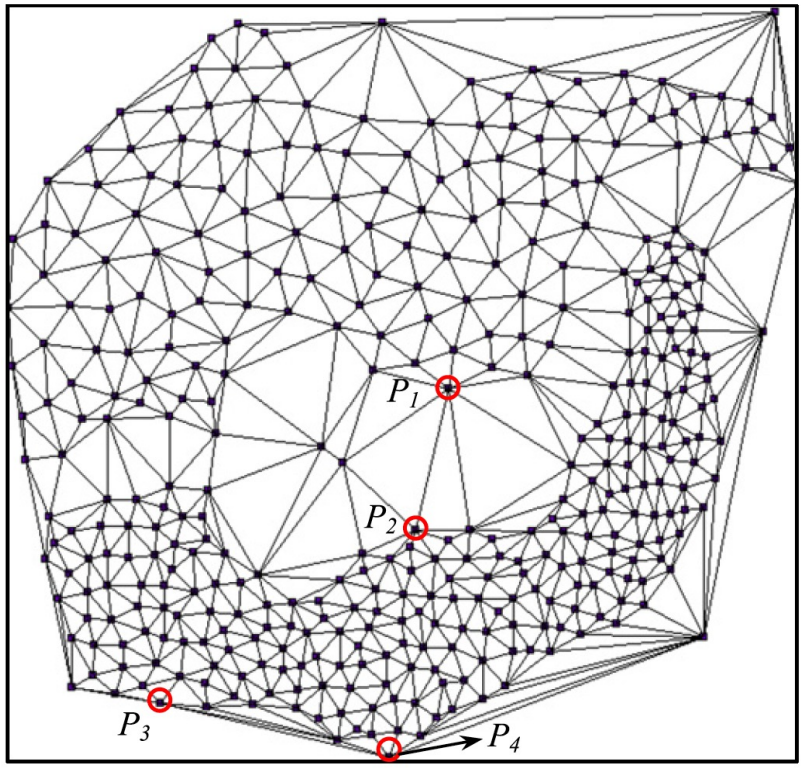
DBSC: An Overview

1. Constructing spatial proximity relationships

Oregon State University
College of Engineering

DBSC: An Overview

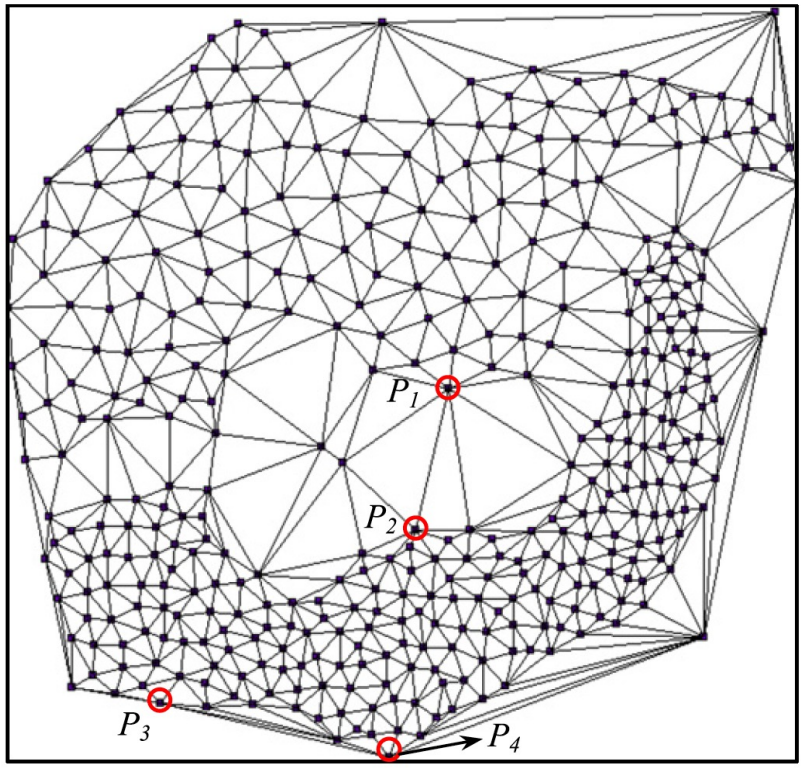
1. Constructing spatial proximity relationships
   a. Delanauy Triangulation



Oregon State University
College of Engineering

DBSC: An Overview

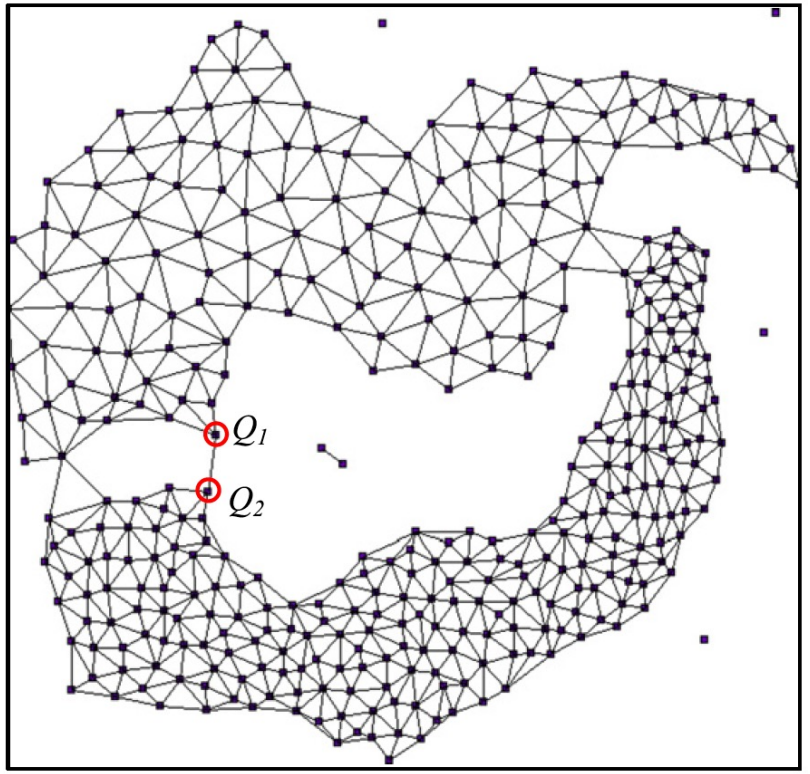
1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges



Oregon State University
College of Engineering

DBSC: An Overview

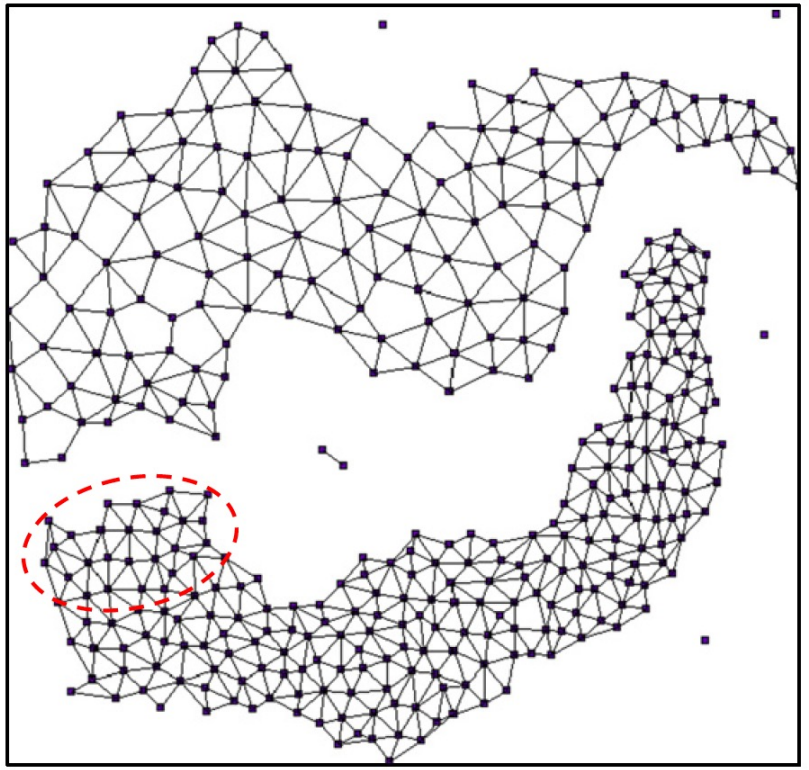
1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges



Oregon State University
College of Engineering

DBSC: An Overview

1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges

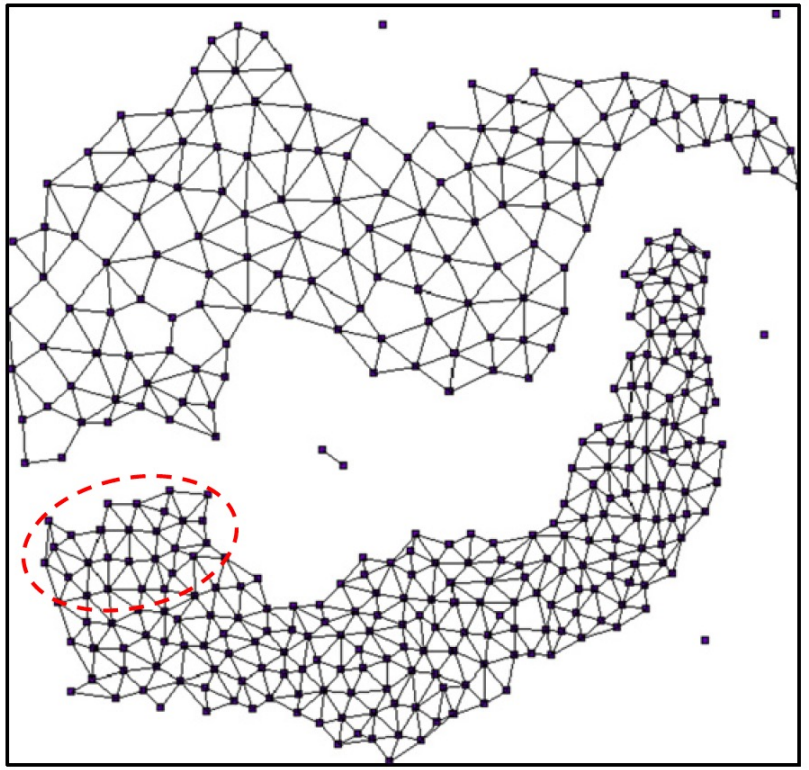


Oregon State University
College of Engineering

DBSC: An Overview

1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges

2. Grouping similar data points

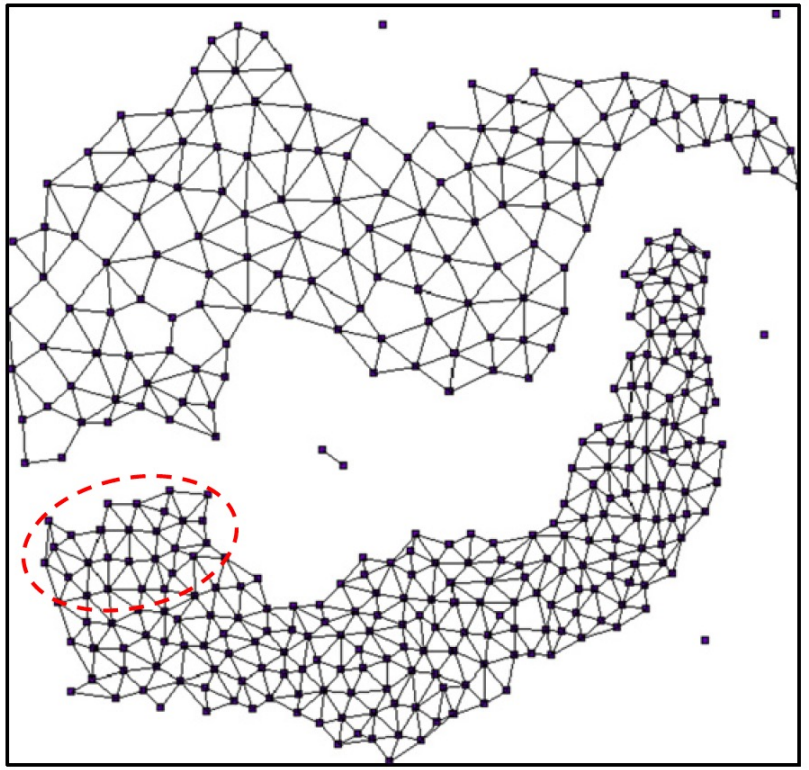


Oregon State University
College of Engineering

DBSC: An Overview

1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges

2. Grouping similar data points
   a. Must be neighbors

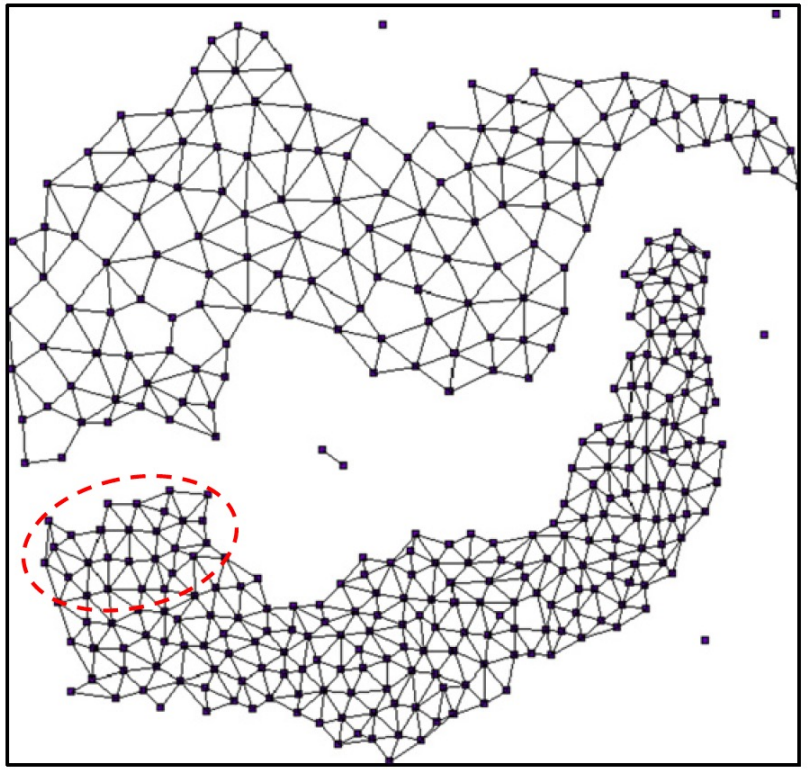


Oregon State University
College of Engineering

DBSC: An Overview

1. Constructing spatial proximity relationships
   a. Delanauy Triangulation
   b. Remove global & local long edges

2. Grouping similar data points
   a. Must be neighbors
   b. Distance in feature space must be below a certain threshold



clustGeo: An Overview

- $d(x,y) = \alpha\, d_1(x,y) + (1-\alpha)\, d_2(x,y)$

Oregon State University
College of Engineering

clustGeo: An Overview

- $d(x,y) = \alpha\, d_1(x,y) + (1-\alpha)\, d_2(x,y)$

spatial distance

environmental distance

Oregon State University
College of Engineering

clustGeo: An Overview

- $d(x,y) = \alpha\, d_1(x,y) + (1-\alpha)\, d_2(x,y)$

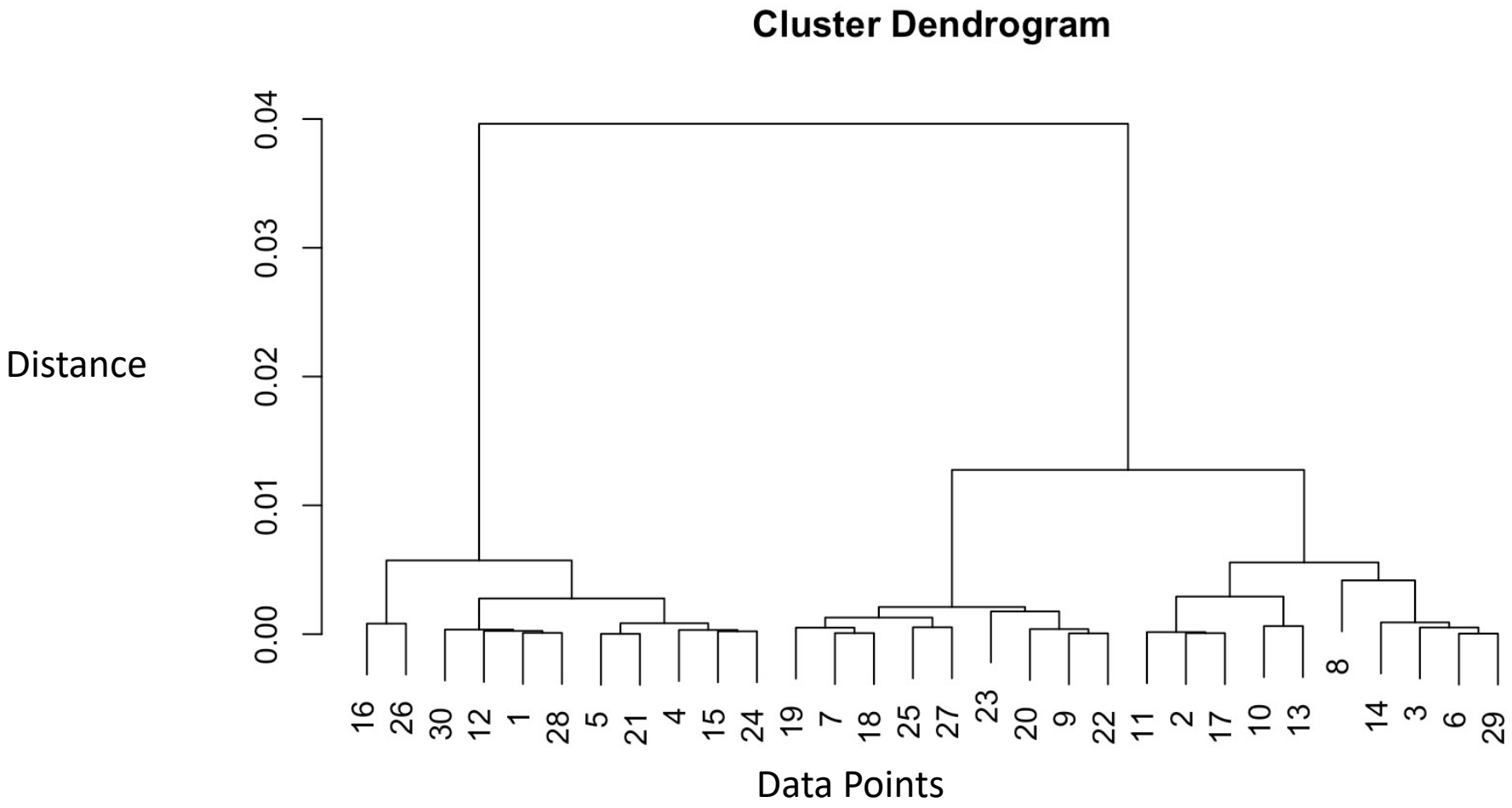
spatial distance

environmental distance

- Continuously cluster the most similar objects together

Oregon State University
College of Engineering

clustGeo: An Overview



Oregon State University
College of Engineering

clustGeo: An Overview

- $\alpha = .8$
- # of sites $\approx$ # of unique locations * .85

Experimental Setup

1. Real Data
2. Semi-Simulated
3. Simulated
4. Adversarial

Oregon State University
College of Engineering

Experimental Setup

1. Real Data
2. Semi-Simulated
3. Simulated
4. Adversarial

} Simulated Species

Oregon State University
College of Engineering

Modeling Process

Oregon State University
College of Engineering

Modeling Process

Clustering

Oregon State University
College of Engineering

Modeling Process

Clustering

OM

Oregon State University
College of Engineering

Modeling Process

Clustering

OM

$$occ\_prob = \beta_0 + \beta_1 * occ\_var_1 + \ldots$$
$$det\_prob = \alpha_0 + \alpha_1 * det\_var_1 + \ldots$$

Oregon State University
College of Engineering

Modeling Process

Clustering

OM

$$occ\_prob = \beta_0 + \beta_1 * occ\_var_1 + \ldots$$
$$det\_prob = \alpha_0 + \alpha_1 * det\_var_1 + \ldots$$

checklist_id	occ_var1	occ_var2	det_var1	det_var2	pred_occ_prob	pred_det_prob
S123	0.25	0.33	0.55	0.37		
S124	0.17	0.38	0.67	0.75		
S125	0.84	0.08	0.15	0.76		

Oregon State University
College of Engineering

Modeling Process

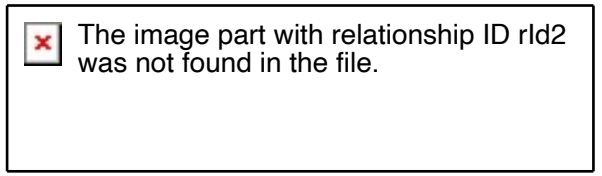
Clustering

OM

$$occ\_prob = \beta_0 + \beta_1 * occ\_var_1 + \ldots$$
$$det\_prob = \alpha_0 + \alpha_1 * det\_var_1 + \ldots$$

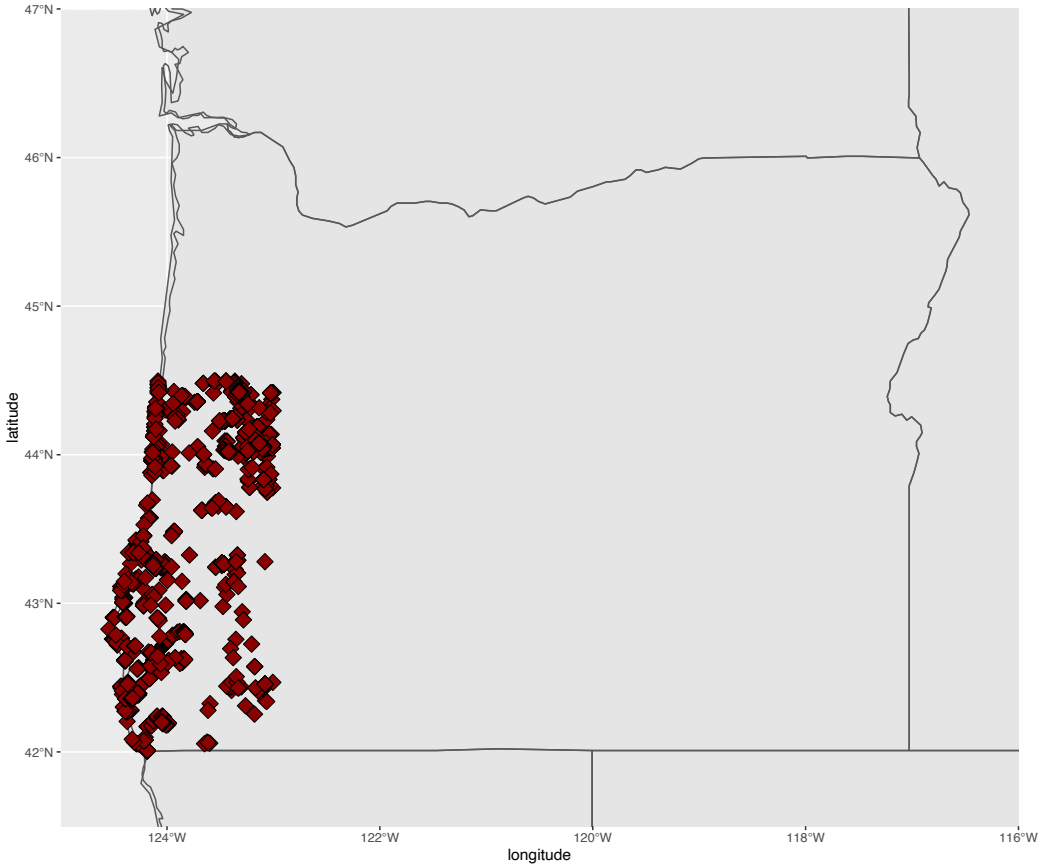
checklist_id	occ_var1	occ_var2	det_var1	det_var2	pred_occ_prob	pred_det_prob
S123	0.25	0.33	0.55	0.37		
S124	0.17	0.38	0.67	0.75		
S125	0.84	0.08	0.15	0.76		

Experiments!



Oregon State University
College of Engineering

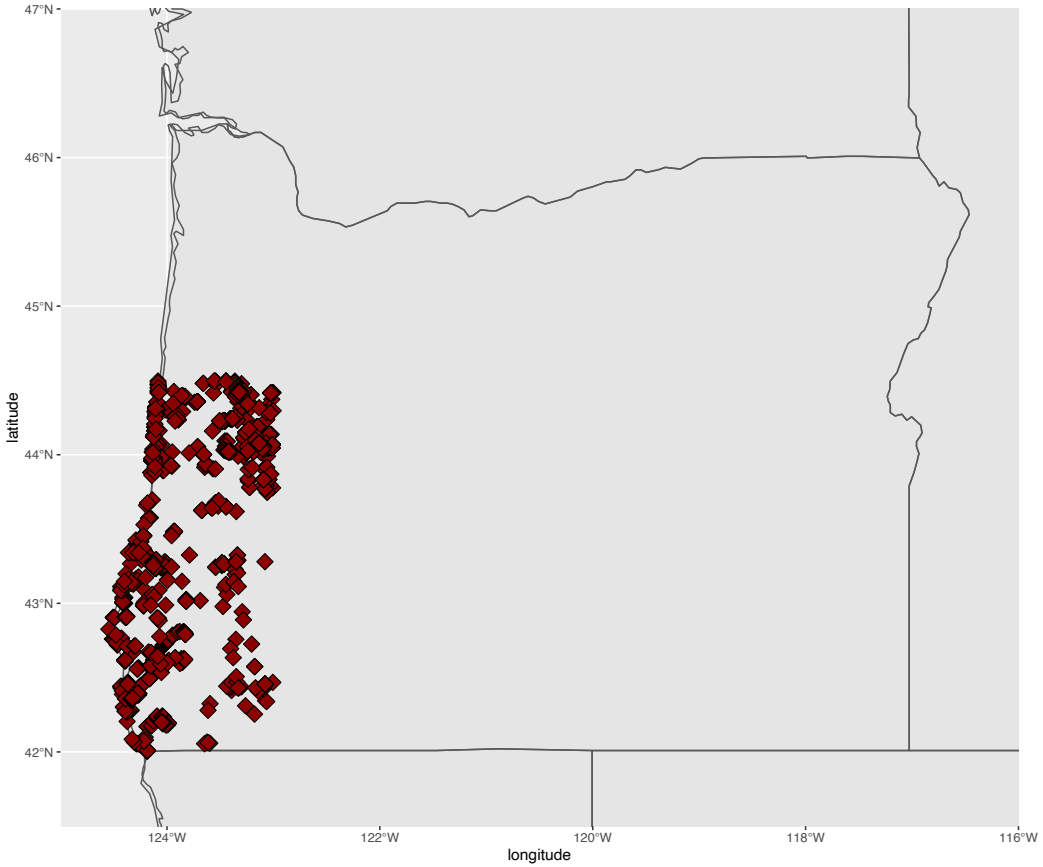
Real Data Experiment



- >2,000 checklists from eBird
  - collected between May and July 2017
  - Occ/Det Variables
  - Detections/Non-detections for WETA & HEWA

Oregon State University
College of Engineering

Real Data Experiment



- >2,000 checklists from eBird
  - collected between May and July 2017
  - Occ/Det Variables
  - Detections/Non-detections for WETA & HEWA

$$occ\_prob = 1.24 + .86 * occ\_var_1 + \ldots$$

$$det\_prob = .99 - .31 * det\_var_1 + \ldots$$

Oregon State University
College of Engineering

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43		1
S124	.	0.17	0.83	0.14		0
S125	.	0.84	0.77	0.65		1

Oregon State University
College of Engineering

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43		1
S124	.	0.17	0.83	0.14		0
S125	.	0.84	0.77	0.65		1

- Discrimination Threshold: what constitutes a *predicted* detection?

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43		1
S124	.	0.17	0.83	0.14		0
S125	.	0.84	0.77	0.65		1

- Discrimination Threshold = .5

Oregon State University
College of Engineering

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	0	1
S124	.	0.17	0.83	0.14	0	0
S125	.	0.84	0.77	0.65	1	1

- Discrimination Threshold = .5

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	0	1
S124	.	0.17	0.83	0.14	0	0
S125	.	0.84	0.77	0.65	1	1

- Discrimination Threshold = .5
  - TPR := TP/P = 1/2
  - FPR := FP/N = 0/1

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	0	1
S124	.	0.17	0.83	0.14	0	0
S125	.	0.84	0.77	0.65	1	1

- Discrimination Threshold = .5
  - TPR := TP/P = 1/2
  - FPR := FP/N = 0/1

Real Data Evaluation: ROC/AUC

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	0	1
S124	.	0.17	0.83	0.14	0	0
S125	.	0.84	0.77	0.65	1	1

- Discrimination Threshold = .5
  - TPR := TP/P = 1/2
  - FPR := FP/N = 0/1

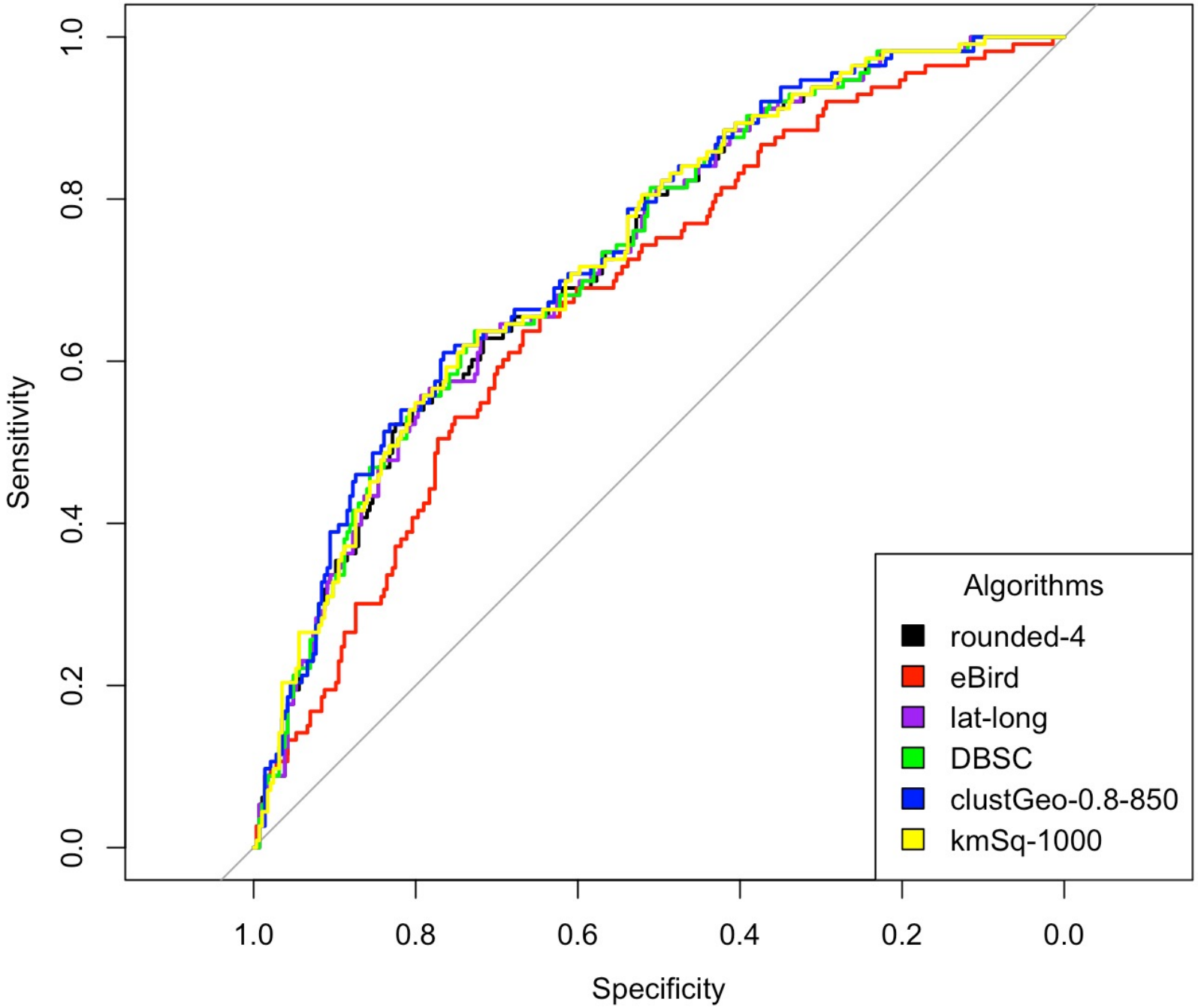
Real Data Evaluation: ROC/AUC

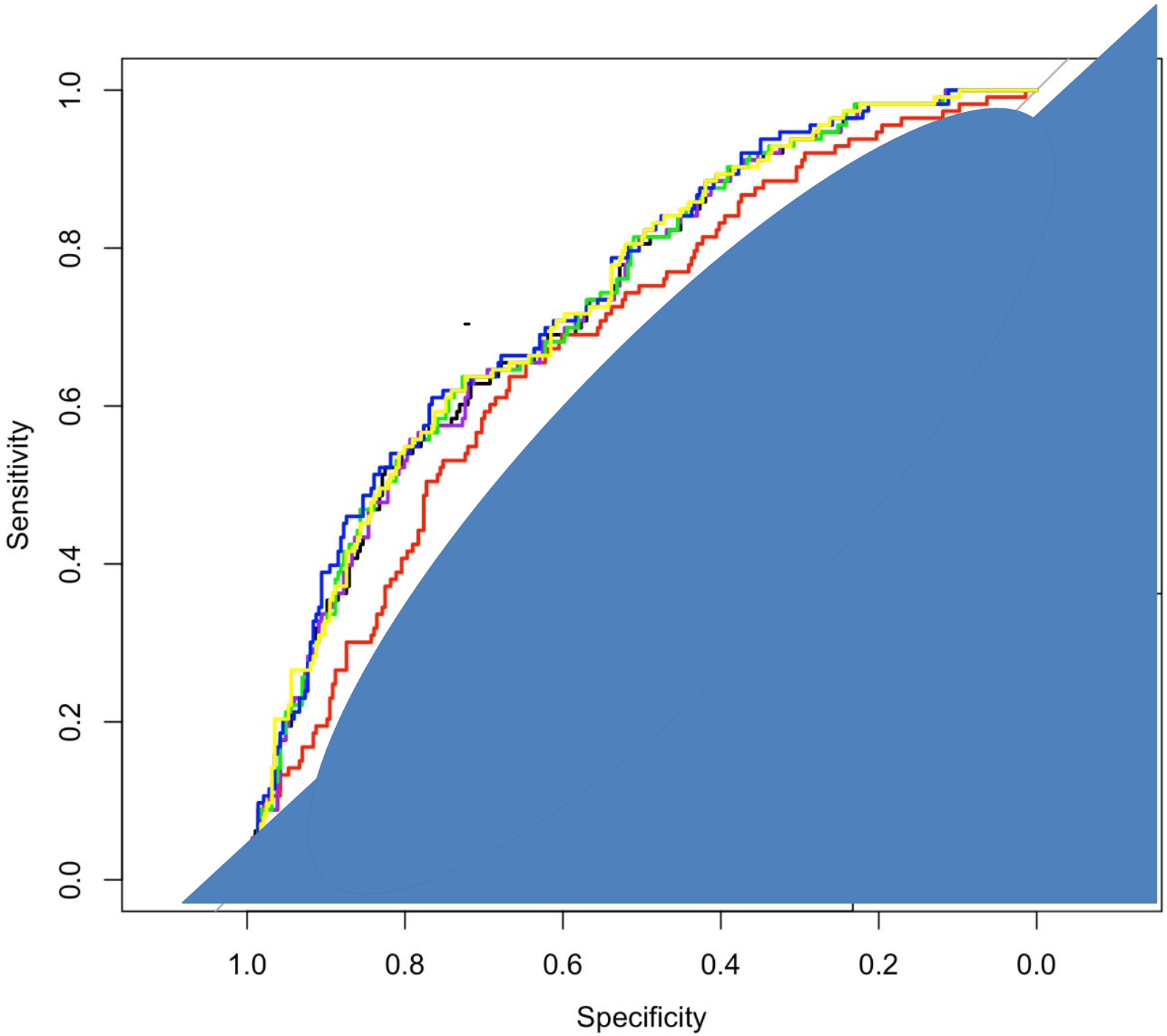
checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	?	1
S124	.	0.17	0.83	0.14	?	0
S125	.	0.84	0.77	0.65	?	1

- What about other discrimination thresholds?

Oregon State University
College of Engineering

ROC Example





Oregon State University
College of Engineering

AUC (Area under the Curve): higher is better

Real Data Evaluation: Training vs Validation

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	?	1
S124	.	0.17	0.83	0.14	?	0
S125	.	0.84	0.77	0.65	?	1

Oregon State University
College of Engineering

Real Data Evaluation: Training vs Validation

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	?	1
S124	.	0.17	0.83	0.14	?	0
S125	.	0.84	0.77	0.65	?	1

1. OR2020: Silver Standard

Oregon State University
College of Engineering

Real Data Evaluation: Training vs Validation

checklist_id	...	occ_prob	det_prob	detected_prob	pred_is_detected	is_detected
S123	.	0.65	0.66	0.43	?	1
S124	.	0.17	0.83	0.14	?	0
S125	.	0.84	0.77	0.65	?	1

1. OR2020: Silver Standard

2. Random Selection

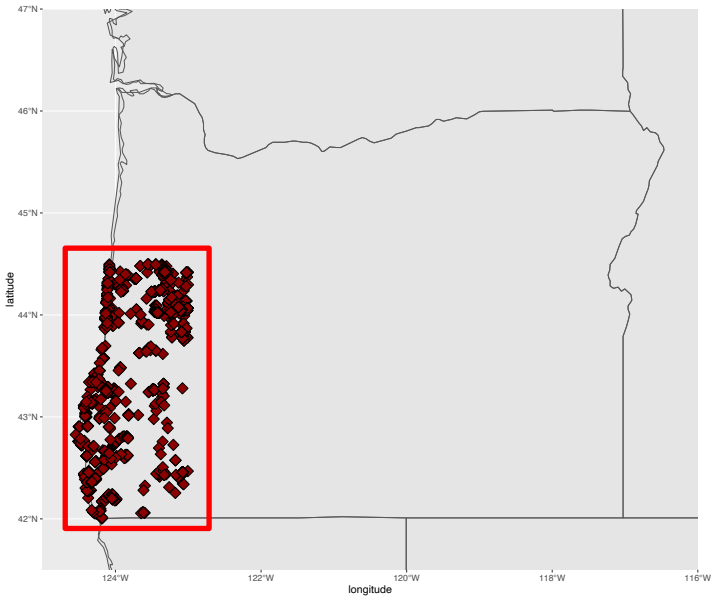
Oregon State University
College of Engineering

Simulated Experiments!

Oregon State University
College of Engineering

Simulated Species Creation

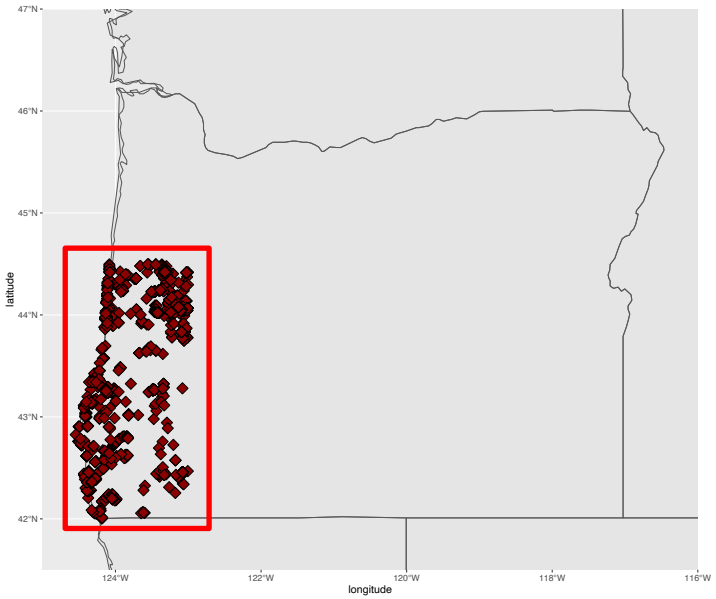
1. Select locations and attach occupancy and detection variables



Oregon State University
College of Engineering

Simulated Species Creation

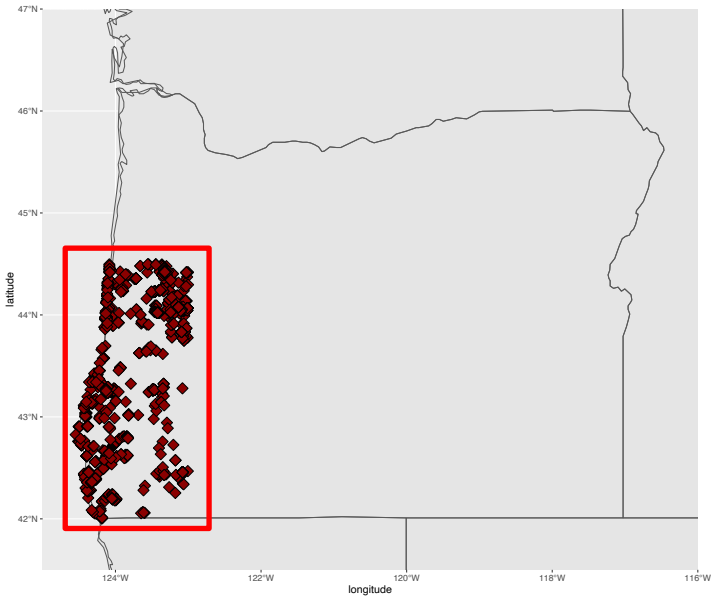
1. Select locations and attach occupancy and detection variables
2. Choose coefficients for generative model
   - $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$



Oregon State University
College of Engineering

Simulated Species Creation

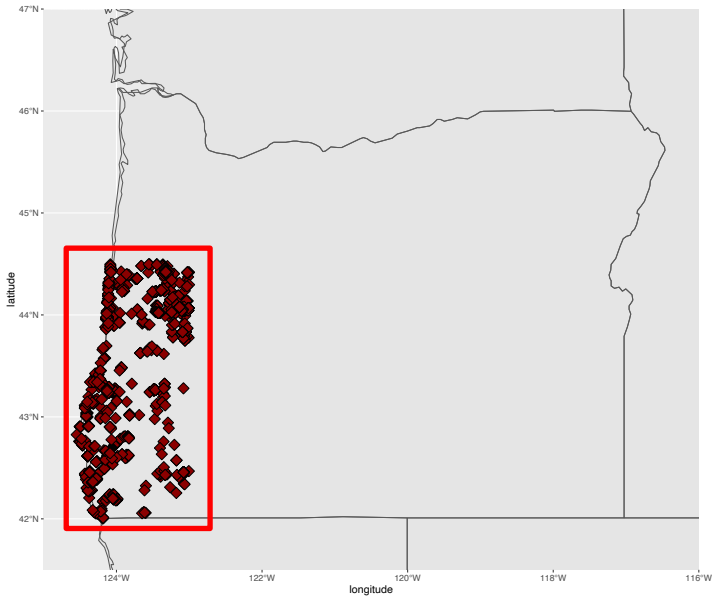
1. Select locations and attach occupancy and detection variables
2. Choose coefficients for generative model
   - $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
3. Calculate occ/det probabilities for each site/checklist



Oregon State University
College of Engineering

Simulated Species Creation

1. Select locations and attach occupancy and detection variables
2. Choose coefficients for generative model
   - $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
3. Calculate occ/det probabilities for each site/checklist

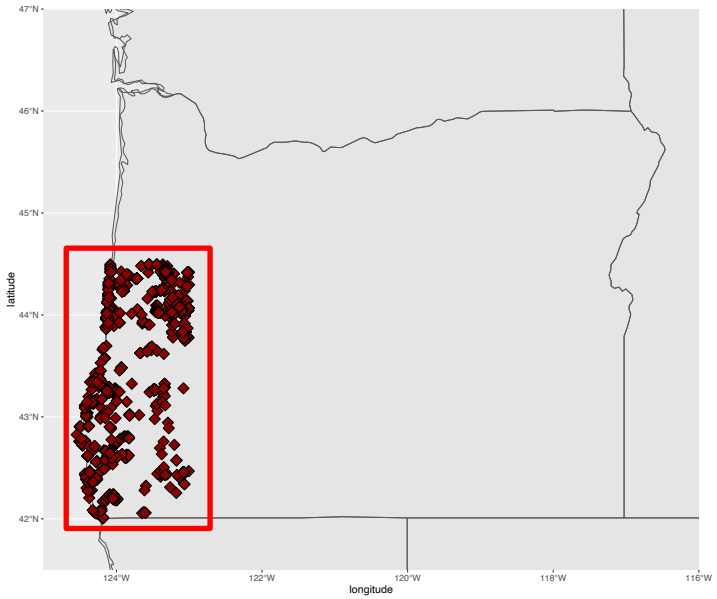


checklist_id	occ_var1	occ_var2	det_var1	det_var2	occ_prob	det_prob
S123	0.25	0.33	0.55	0.37	0.62	0.62
S124	0.17	0.38	0.67	0.75	0.51	0.74
S125	0.84	0.08	0.15	0.76	0.86	0.36

Oregon State University
College of Engineering

Simulated Species Creation

1. Select locations and attach occupancy and detection variables
2. Choose coefficients for generative model
   - $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
3. Calculate occ/det probabilities for each site/checklist



checklist_id	occ_var1	occ_var2	det_var1	det_var2	occ_prob	det_prob	is_occupied	is_detected
S123	0.25	0.33	0.55	0.37	0.62	0.62	1	0
S124	0.17	0.38	0.67	0.75	0.51	0.74	0	0
S125	0.84	0.08	0.15	0.76	0.86	0.36		

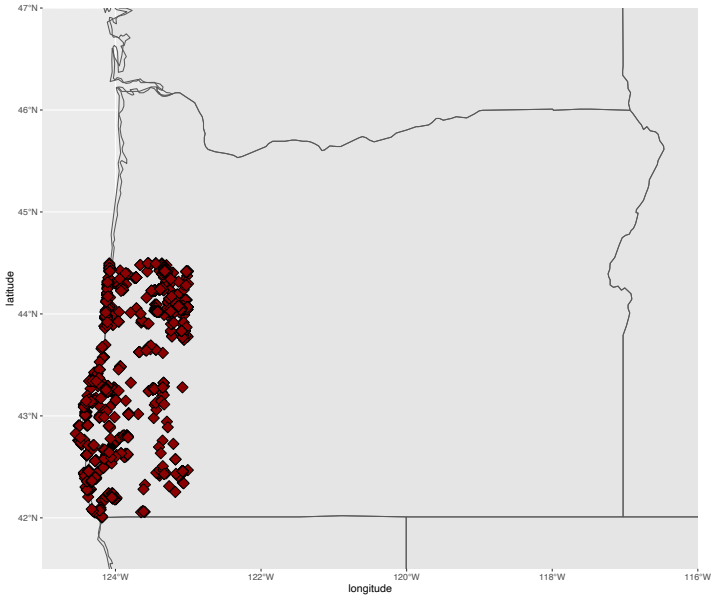
Oregon State University
College of Engineering

On to the experiments…

Oregon State University
College of Engineering

Semi-Simulated Experiments

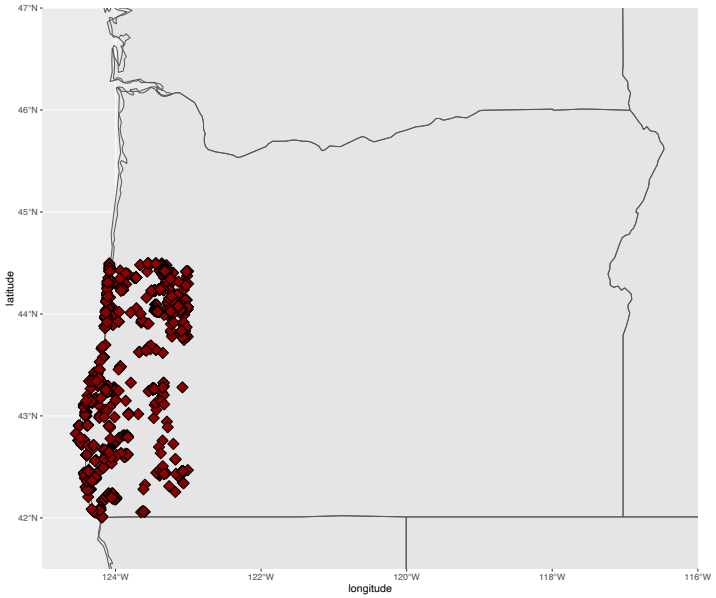
- Derived from eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$



Oregon State University
College of Engineering

Semi-Simulated Experiments

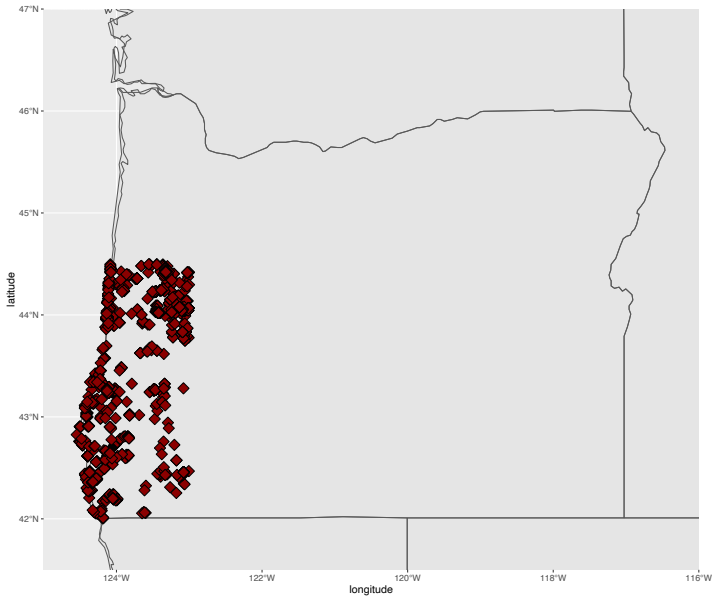
- Derived from eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
- Evaluation
  - Compare occ/det probabilities
  - Cluster similarity



Oregon State University
College of Engineering

Semi-Simulated Experiments

- Derived from eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
- Evaluation
  - Compare occ/det probabilities
  - Cluster similarity

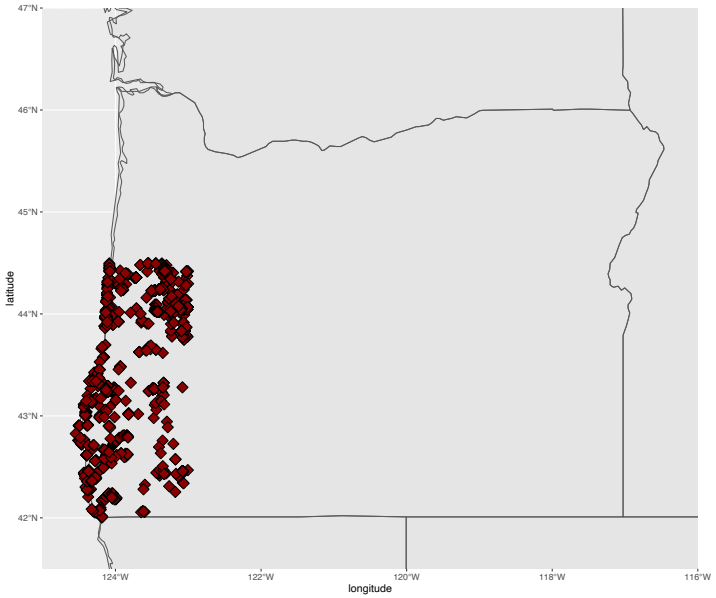


checklist_id	occ_var1	occ_var2	det_var1	det_var2	occ_prob	pred_occ_prob	det_prob	pred_det_prob
S123	0.25	0.33	0.25	0.33	0.62	0.54	0.62	0.61
S124	0.17	0.38	0.17	0.38	0.51	0.24	0.74	0.66
S125	0.84	0.08	0.84	0.08	0.86	0.61	0.36	0.49

Oregon State University
College of Engineering

Semi-Simulated Experiments

- Derived from eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$



- Evaluation
  - Compare occ/det probabilities
  - Cluster similarity

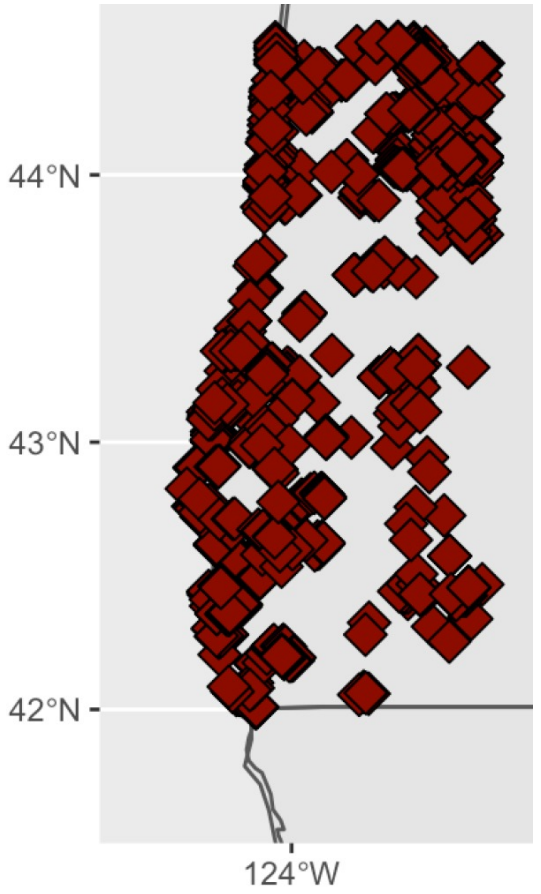
Mean Squared Error (MSE)

checklist_id	occ_var1	occ_var2	det_var1	det_var2	occ_prob	pred_occ_prob	det_prob	pred_det_prob
S123	0.25	0.33	0.25	0.33	0.62	0.54	0.62	0.61
S124	0.17	0.38	0.17	0.38	0.51	0.24	0.74	0.66
S125	0.84	0.08	0.84	0.08	0.86	0.61	0.36	0.49

Oregon State University
College of Engineering

Cluster Similarity

- Constructed a ground truth clustering



Oregon State University
College of Engineering

Cluster Similarity

- Constructed a ground truth clustering

- Cluster similarity
  - Adjusted Rand Index (ARI)
  - Adjusted Mutual Index (AMI)
  - Normalized Information Distance (NID) (Vinh et al., 2010)

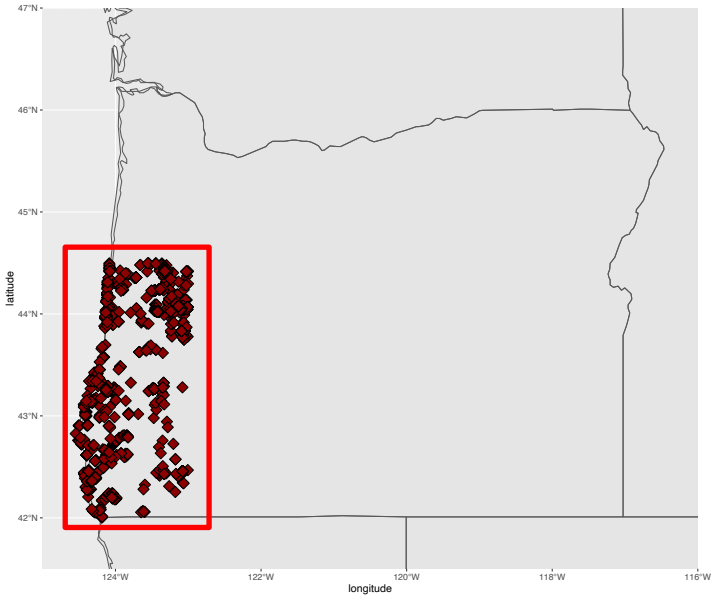
Ground Truth Clustering

Clustering 1

Oregon State University
College of Engineering

Simulated Experiments

- Chosen at random from the same region as the eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$



Oregon State University
College of Engineering

Simulated Experiments

- Chosen at random from the same region as the eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$

Oregon State University
College of Engineering

Simulated Experiments

- Chosen at random from the same region as the eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
- Varied 4 attributes, 1 at a time:
  - Occupancy & detection probability
  - # of unique locations
  - Average # of visits to a location

Oregon State University
College of Engineering

Simulated Experiments

- Chosen at random from the same region as the eBird checklists
  - Define $\text{occ prob} = o\_var_0 + .75 * o\_var_1 - 1.25 * o\_var_2$
- Varied 4 attributes, 1 at a time:
  - Occupancy & detection probability
  - # of unique locations
  - Average # of visits to a location
- Evaluation
  - Compare occ/det probabilities

Results: Real-Data Experiments

Oregon State University
College of Engineering

WETA AUC

method	Random	OR2020
eBird-BP	.7300 ± .027	.5620
1-kmSq	.7333 ± .022	.5328
lat-long	.7387 ± .022	.5438
rounded-4	.7387 ± .022	.5444
DBSC	.7308 ± .022	.5367
clustGeo	.7389 ± .022	.5550

- *Both validation sets have 65 1s & 334 0s*
- *Prevalence of…*
  - *Validation sets 16.3%*
  - *Training sets 11.8%*

HEWA AUC

method	Random	OR2020
eBird-BP	.6926 ± .117	.5891
1-kmSq	.8677 ± .021	.7225
lat-long	.8675 ± .021	.7289
rounded-4	.8676 ± .021	.7289
DBSC	.8661 ± .018	.7300
clustGeo	.8684 ± .020	.7408

- *Both validation sets have 113 1s & 286 0s*
- *Prevalence of…*
  - *Validation sets 28.3%*
  - *Training sets 8.9%*

Results : Semi-Simulated Experiments

Oregon State University
College of Engineering

	ARI	AMI	NID	occ MSE
ground truth	1.0	1.0	0	.0389 ± .015
eBird-BP	-	-	-	.1177 ± .041
1-kmSq	.9948	.9401	.0599	.1065 ± .027
lat-long	.9992	.9825	.0175	.0422 ± .017
rounded-4	.9992	.9826	.0174	.0424 ± .017
DBSC	.9806	.9566	.0434	.1193 ± .031
clustGeo	.9994	.9909	.0091	.0460 ± .019

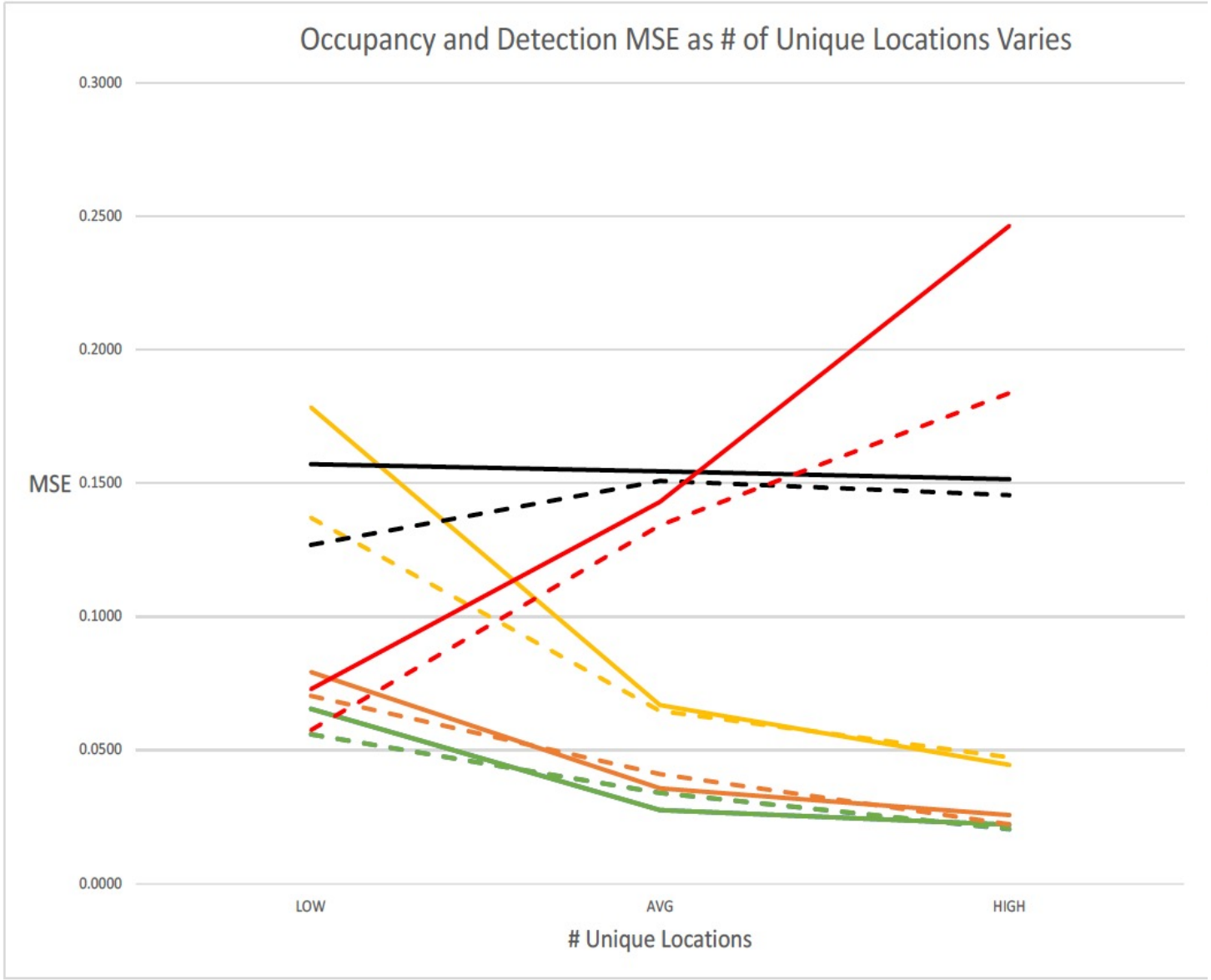
Results : Semi-Simulated Experiments

Oregon State University
College of Engineering

	ARI	AMI	NID	occ MSE
ground truth	1.0	1.0	0	.0389 ± .015
eBird-BP	-	-	-	.1177 ± .041
1-kmSq	.9948	.9401	.0599	.1065 ± .027
lat-long	.9992	.9825	.0175	.0422 ± .017
rounded-4	.9992	.9826	.0174	.0424 ± .017
DBSC	.9806	.9566	.0434	.1193 ± .031
clustGeo	.9994	.9909	.0091	.0460 ± .019

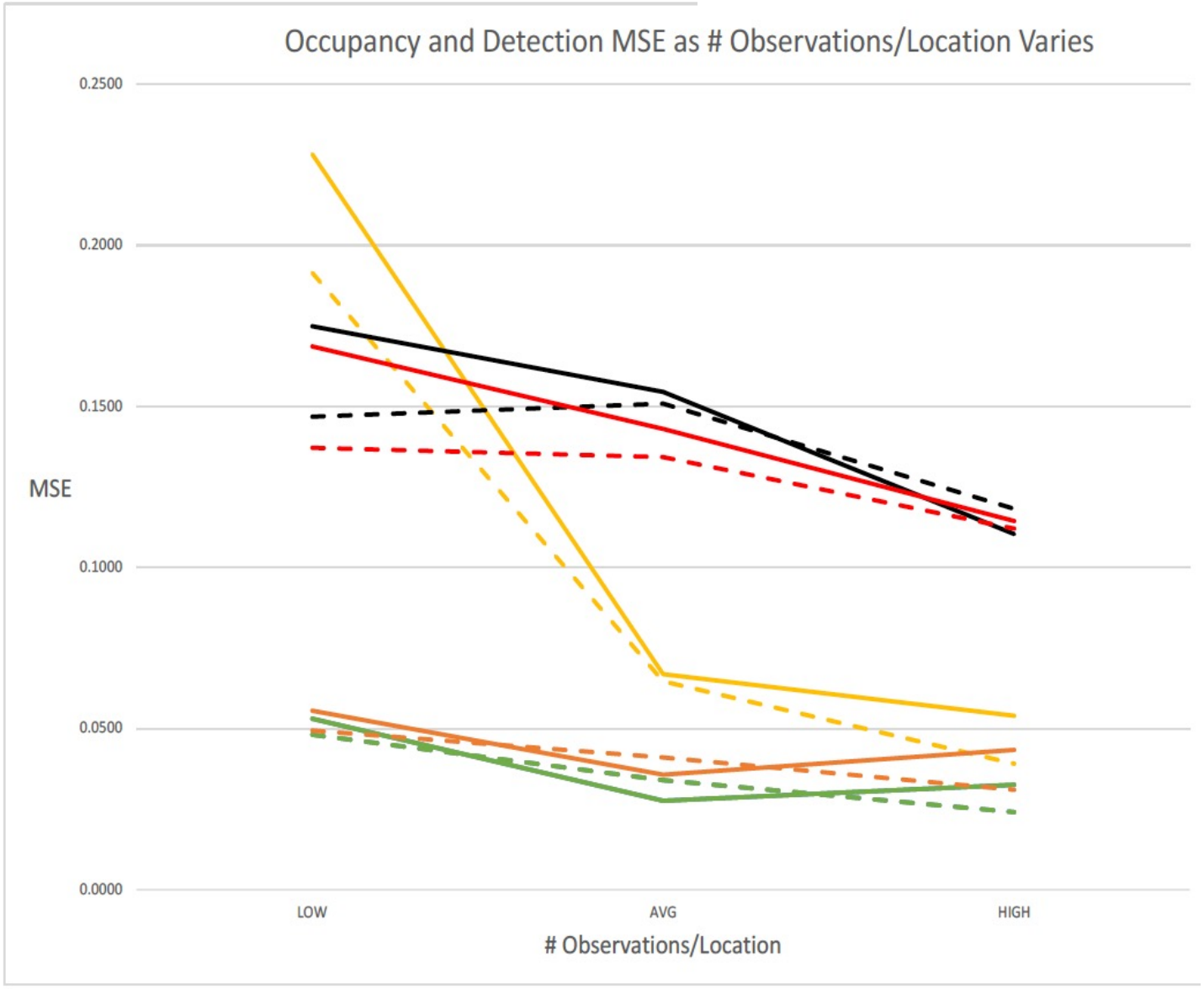
Results : Simulated Experiments

Oregon State University
College of Engineering



Results : Simulated Experiments

Oregon State University
College of Engineering



Oregon State University
College of Engineering

Contributions

1. Introduced the **Site Clustering Problem**

2. Identified solutions that outperform the existing approaches

3. Learned ways that algorithms struggle to address this challenge

Great Egret, eBird.org

Oregon State University
College of Engineering

Impact

1. Increases accuracy of (avian) SDMs built from community science data
   - Informs conservation decisions to mitigate (avian) biodiversity loss

Oregon State University
College of Engineering

Impact

1. Increases accuracy of (avian) SDMs built from community science data
   - Informs conservation decisions to mitigate (avian) biodiversity loss

2. Introduces a framework to model species at spatial & temporal scales previously unattainable

Oregon State University
College of Engineering

Impact

1. Increases accuracy of (avian) SDMs built from community science data
   - Informs conservation decisions to mitigate (avian) biodiversity loss

2. Introduces a framework to model species at spatial & temporal scales previously unattainable

3. Builds upon existing literature that validates community science data

Oregon State University
College of Engineering

Future Work

Oregon State University
College of Engineering

Future Work

- Further investigate consensus clustering

Consensus Clustering

Oregon State University
College of Engineering

Clustering 1

Consensus Clustering

Oregon State University
College of Engineering

Clustering 1

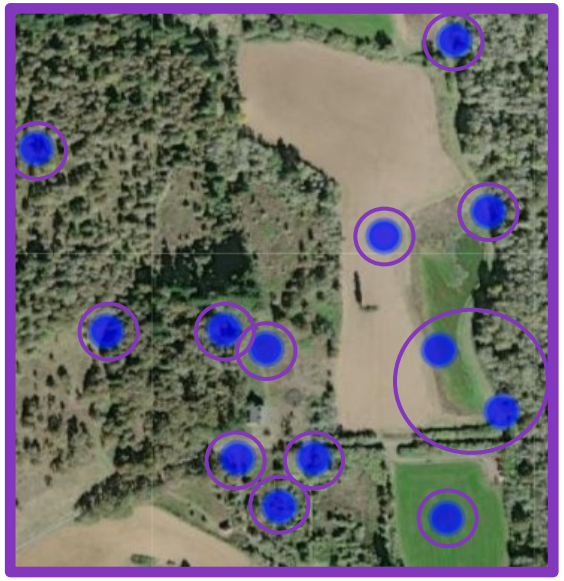
Clustering 2

Consensus Clustering

Oregon State University
College of Engineering

Clustering 1

Clustering 2



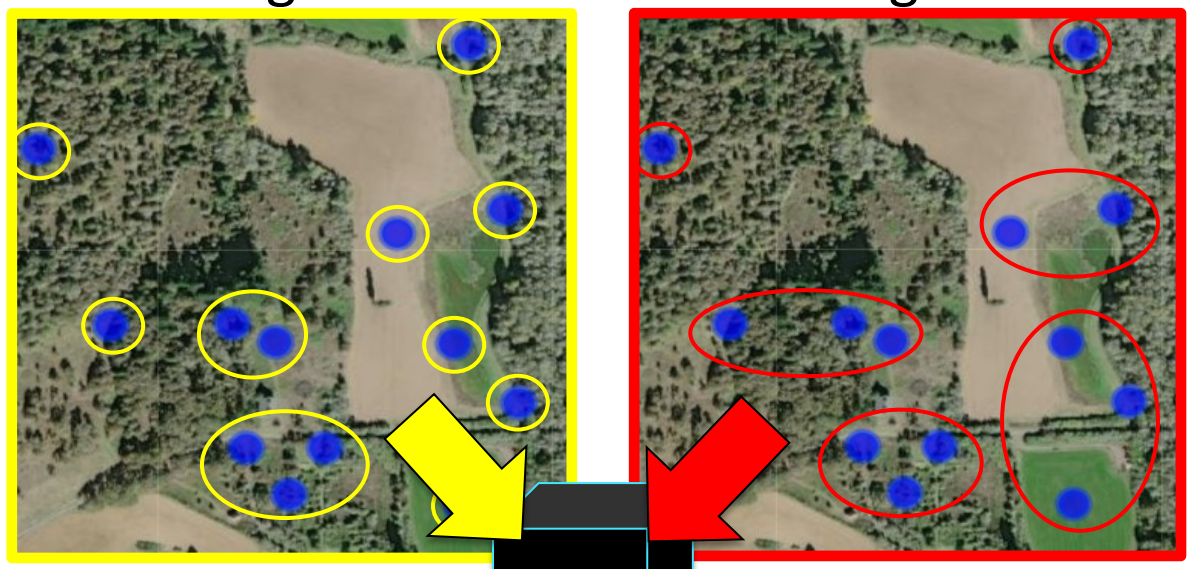
Clustering i

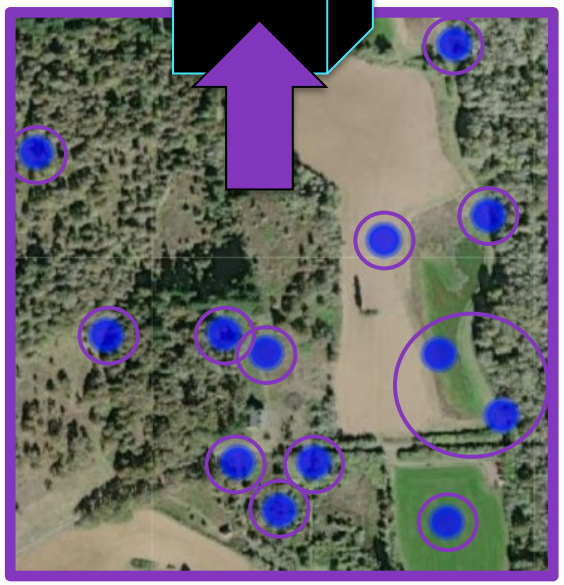
Consensus Clustering

Oregon State University
College of Engineering

Clustering 1

Clustering 2





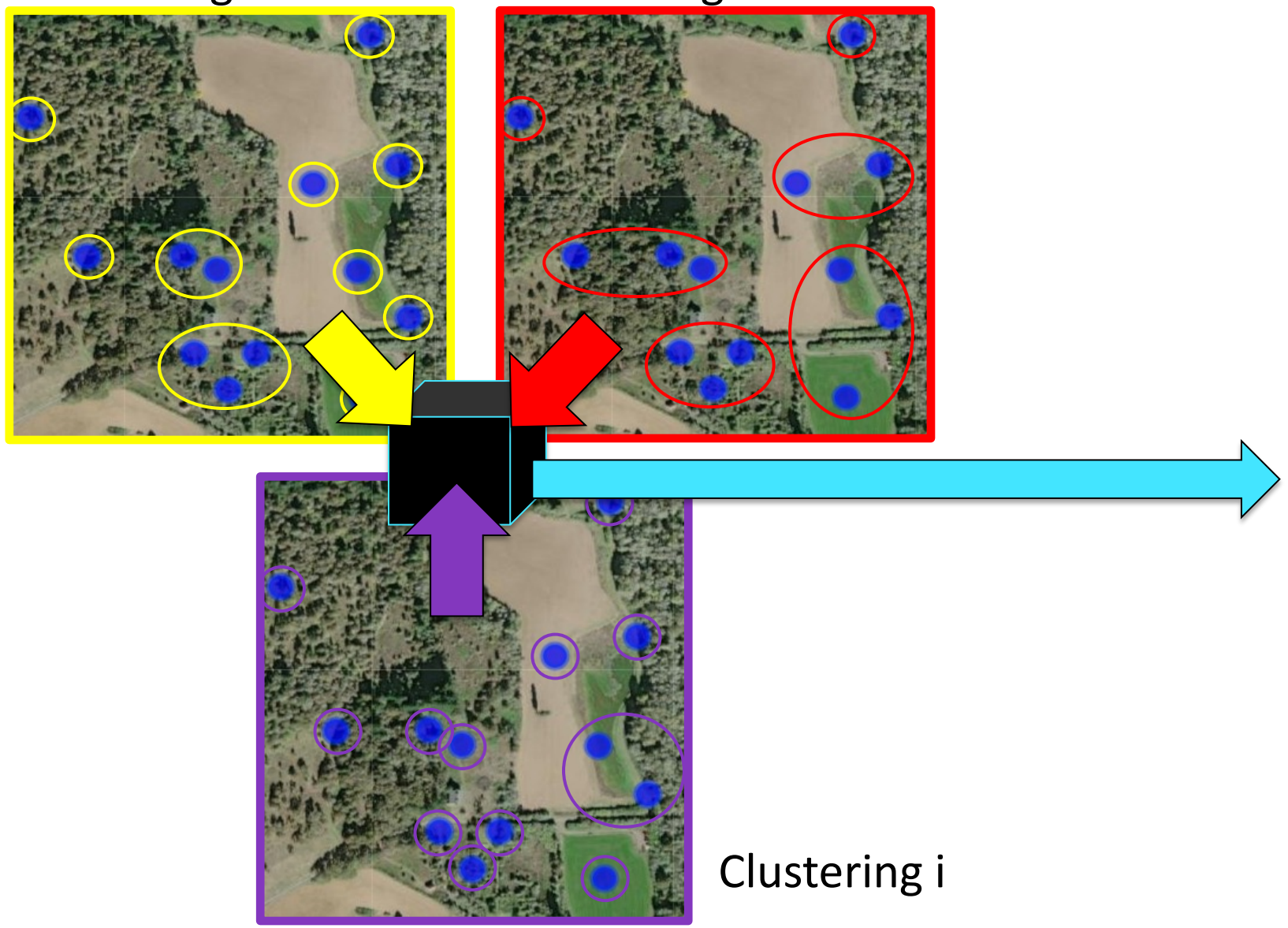
Clustering i

Consensus Clustering

Oregon State University
College of Engineering

Clustering 1

Clustering 2



Clustering i

Consensus Clustering Result

Oregon State University
College of Engineering

Future Work

- Further investigate consensus clustering
- Domain-informed spatial clustering algorithm
  - Home range
  - Bird call strength (aka, maximum detection distance)

Oregon State University
College of Engineering

Future Work

- Further investigate consensus clustering
- Domain-informed spatial clustering algorithm
  - Home range
  - Bird call strength (aka, maximum detection distance)
- Extend solutions to be spatio-temporal

Oregon State University
College of Engineering

Future Work

- Further investigate consensus clustering
- Domain-informed spatial clustering algorithm
  - Home range
  - Bird call strength (aka, maximum detection distance)
- Extend solutions to be spatio-temporal
- Transfer learnings to similar settings

References

Assunção, R. M., Neves, M. C., Câmara, G., and Freitas, C. D. C. Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7):797–811, 2006.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. Clustgeo: an r package for hierarchical clustering with spatial constraints. Computational Statistics, 33(4): 1799–1822, Jan 2018.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A densitybased algorithm for discovering clusters in large spatial databases with noise. pp. 226–231, 1996.

Gionis, A., Mannila, H., and Tsaparas, P. Clustering aggregation. ACM Trans. Knowl. Discov. Data, 1(1):4–es, March 2007. ISSN 1556-4681.

Guillera-Arroita, G., Lahoz-Monfort, J., MacKenzie, D. I., Wintle, B. A., and McCarthy, M. A. Ignoring imperfect detection in biological surveys is dangerous: a response to 'fitting and interpreting occupancy models'. PloS one, 9(7):e99571–e99571, 07 2014.

Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). International Journal of Geographical Information Science, 22 (7):801–823, 2008.

Johnston, Alison, et al. "Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions." Diversity and Distributions (2021).

Liu, Q., Deng, M., Shi, Y., and Wang, J. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers Geosciences, 46:296–309, 2012.

Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

Ng, R. and Han, J. Clarans: a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, 14(5):1003–1016, 2002.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res., 11:2837–2854, December 2010. ISSN 1532-4435.

Oregon State University
College of Engineering

Oregon State University
College of Engineering

Thank You!

- To my advisor: Dr. Rebecca Hutchinson
- To my committee: Dr. William Douglas Robinson, Dr. Weng-Keen Wong
- To collaborators: Dr. Tyler Hallman, Jack Kilbride
- To my research group: Dr. Eugene Seo, Laurel Hopkins, Jing Wang, Vishnu Reghunathan, Nahian Ahmed, Andrew Droubay, Chelsea Li, Nate Butler, Demetrius Hernandez, AnaPatricia Medina
- To my family, friends, and extended support network

Oregon State University
College of Engineering

Questions?

https://i.pinimg.com/736x/57/4f/92/574f92c2442eebd20a158fd4e2c013d9--human-emotions-owl-eyes.jpg